

Improving the Design and Usability of Password Creation Systems

Saja Abdullah Ayed Althubaiti

Doctor of Philosophy

University of York

Computer Science

January 2018

To my family, beloved husband, and Azzam

Abstract

The aim of this research is to inform the design and usability of password creation systems (PCSs) and their supporting features so they can better support users when creating passwords. PCSs are a particular class of an interactive system that allow users to create passwords and which may offer supporting features to help users that process. The supporting features include statements of password policy, password creation suggestions, and password strength indicators. The thesis addresses this aim at the user interface level by providing knowledge about how users react to a range of aspects of the supporting features in PCSs and by providing a set of usability heuristics and guidelines to support the evaluation and development of existing PCSs. There were three phases of research, each consisting of studies that provided insights for the next one. The first phase focused on understanding current practices in PCSs and their effects on users. The outcome of this phase revealed a high number of usability problems including lack of supporting features, design presentation flaws, and ambiguity in password instructions. The second phase investigated the effects of different design aspects in PCSs. The findings showed that different design aspects of PCSs had a significant effect on the usability and password strength in different ways. The third phase proposed a set of usability heuristics and guidelines specifically for the evaluation and design of PCSs. The heuristics were evaluated by usability professionals and were perceived to be easy to understand, clear, and useful. A mixture of qualitative and quantitative methods was used to answer the research questions. The findings suggest that PCSs can effectively support users in creating passwords by addressing four key factors: (1) provision of supporting features, (2) user instructions for creating passwords, (3) timing of presentation for presenting statements of policy and creation suggestions, and (4) media and colour scheme for designing strength indicators.

Contents

Abstract	v
Contents	vii
List of Tables	xiii
List of Figures	xxi
Acknowledgments	xxv
Declaration	xxvii
Chapter 1 Introduction	1
1.1 Research Motivation and Aims	3
1.2 Research Approach and Methodology	5
1.2.1 Phase 1: Understanding the Current Practices of PCSs	6
1.2.2 Phase 2: Testing Design Variables for PCSs	8
1.2.3 Phase 3: Proposing Usability Heuristics and Guidelines for PCSs....	10
1.3 Research Validity	11
1.4 Research Scope	12
1.5 Research Contributions	14
1.6 Statement of Ethics	15
1.7 Thesis Structure.....	16
Chapter 2 Literature Review	17
2.1 Introduction	17
2.2 User Authentication	17
2.2.1 Definition of User Authentication.....	17
2.2.2 Knowledge-Based Authentication Mechanism.....	20
2.2.3 Token-Based Authentication Mechanism	21
2.2.4 Biometric-Based Authentication Mechanism	23
2.3 Knowledge-Based Authentication: Textual Passwords	24
2.3.1 Definition of Textual Passwords	25
2.3.2 Problems with Textual Passwords	25
2.4 Password Creation Systems	26

2.4.1	Password Choice	27
2.4.2	Supporting Features for Creating Passwords	29
2.4.3	Designing Password Creation Systems	34
2.5	User Behaviours Related to Password Creation and Management	35
2.5.1	Problems with Passwords	35
2.5.2	Coping Strategies	37
2.6	Password Security	40
2.6.1	Password Strength	40
2.6.2	Security Threats	40
2.7	Conclusions	42
	<i>Phase 1: Understanding the Current Practices of PCSs</i>	43
Chapter 3	An Analysis of 30 Current PCSs – <i>Study 1</i>	45
3.1	Introduction	45
3.2	Method	46
3.3	Results	47
3.3.1	Components of the 30 PCSs	47
3.3.2	Structure of the 30 PCSs	54
3.4	Discussion	58
3.4.1	Components of Current PCSs	58
3.4.2	Structure of Current PCSs	59
3.5	Conclusions	60
Chapter 4	An Expert Evaluation of 12 Current PCSs – <i>Study 2</i>	61
4.1	Introduction	61
4.2	Method	63
4.2.1	Design	63
4.2.2	Experts	63
4.2.3	PCSs	65
4.2.4	Equipment and Materials	65
4.2.5	Pilot of the Study Procedure	66
4.2.6	Procedure	66
4.2.7	Data Analysis	67
4.3	Results	68

4.3.1	Expert Agreement	68
4.3.2	Usability Problems	70
4.3.3	Categorisation of Usability Problems	71
4.4	Discussion	75
4.5	Conclusions	76
Chapter 5	A User Evaluation of Six Current PCSs – <i>Study 3</i>	77
5.1	Introduction	77
5.2	Method	78
5.2.1	Design	78
5.2.2	Participants	78
5.2.3	PCSs	78
5.2.4	Equipment and Materials	80
5.2.5	Pilot of the Study Procedure	80
5.2.6	Procedure	80
5.2.7	Data Analysis	82
5.3	Results	82
5.3.1	Usability Problems	82
5.3.2	Categorisation of Usability Problems	84
5.3.3	Participant Perceptions of the Six PCSs	88
5.3.4	Comparison of Expert and User Evaluations	91
5.4	Discussion	96
5.5	Conclusions	98
Chapter 6	The Effects of Current PCS Practices on Password Creation and Recall – <i>Study 4</i>	99
6.1	Introduction	99
6.2	Method	99
6.2.1	Design	99
6.2.2	Participants	101
6.2.3	Materials	102
6.2.4	Pilot of the Study Procedure	112
6.2.5	Procedure	112
6.2.6	Data Analysis	113
6.3	Results	113

6.3.1	Password Creation	113
6.3.2	Password Recall	125
6.3.3	Users' Common Password Creation and Recall Practices.....	127
6.4	Discussion	129
6.5	Conclusions.....	132
<i>Phase 2: Testing Design Variables for PCSs</i>		133
Chapter 7 Instructions for Creating Passwords: Analysis and User Study – <i>Study 5</i>		135
7.1	Introduction.....	135
7.2	Analysis of Current Instructions for Creating Passwords.....	137
7.2.1	Data Sources and Coding Scheme	137
7.2.2	Results: Current State of Instructions for Creating Passwords.....	140
7.3	User Study on Instructions for Creating Passwords	144
7.3.1	Method	145
7.3.2	Results.....	157
7.4	Discussion.....	177
7.4.1	Instructions at the Before-Interaction Step	177
7.4.2	Instructions at the During-Interaction Step.....	179
7.4.3	Instructions at the After-Interaction Step.....	181
7.5	Conclusions.....	182
Chapter 8 The Individual Effects of Supporting Features on Password Creation and Recall – <i>Study 6</i>		185
8.1	Introduction.....	185
8.2	Method	187
8.2.1	Design	187
8.2.2	Participants.....	190
8.2.3	Materials	192
8.2.4	Pilot of the Study Procedure	200
8.2.5	Procedure	201
8.2.6	Data Analysis	201
8.3	Results.....	202
8.3.1	Password Policy	203
8.3.2	Password Creation Suggestion.....	214

8.3.3	Password Strength indicator	225
8.3.4	Users' Common Password Creation and Recall Practices	241
8.4	Discussion	242
8.4.1	Password Policy	243
8.4.2	Password Creation Suggestion	246
8.4.3	Password Strength Indicator	247
8.5	Conclusions	249
Chapter 9	The Combined Effects of Supporting Features on Password Creation and Recall – <i>Study 7</i>	251
9.1	Introduction	251
9.2	Method	252
9.2.1	Design	252
9.2.2	Participants	254
9.2.3	Materials	256
9.2.4	Pilot of the Study Procedure	262
9.2.5	Procedure	262
9.2.6	Data Analysis	262
9.3	Results	263
9.3.1	Password Creation	263
9.3.2	Password Recall	275
9.3.3	Comparison Between Individual and Combined Effects	278
9.3.4	Users' Common Password Creation and Recall Practices	298
9.4	Discussion	300
9.5	Conclusions	304
	<i>Phase 3: Proposing Usability Heuristics and Guidelines for PCSs</i>	305
Chapter 10	Password Creation System Heuristics and Guidelines: Development and Evaluation - <i>Study 8</i>	307
10.1	Introduction	307
10.2	Development of the <i>PassHeuristics</i> and <i>PassGuidelines</i>	309
10.2.1	Data Sources	309
10.2.2	First Version of the <i>PassHeuristics</i> and <i>PassGuidelines</i>	311
10.2.3	Review Process of the First Version of the <i>PassHeuristics</i> and <i>PassGuidelines</i>	324

10.2.4	Final Version of the <i>PassHeuristics</i> and <i>PassGuidelines</i>	331
10.3	Evaluation of the <i>PassHeuristics</i>	333
10.3.1	Method	333
10.3.2	Results	339
10.4	Discussion	341
10.5	Conclusions	344
Chapter 11	General Discussion and Conclusions	345
11.1	Overall Summary	346
11.1.1	Phase 1: Understanding the Current Practices of PCSs	346
11.1.2	Phase 2: Testing Design Variables for PCSs	347
11.1.3	Phase 3: Proposing Usability Heuristics and Guidelines for PCSs..	348
11.2	Discussion	348
11.2.1	Implications from the System Side Perspective	349
11.2.2	Implications from the User Side Perspective	354
11.3	Password-Related Studies	355
11.4	Limitations and Future Work	357
11.5	Conclusions	358
References	361

List of Tables

Table 3.1 PCSs analysed for this study	46
Table 3.2 Frequency of the two language attributes used in the password policy ($N = 27$)	49
Table 3.3 Frequency of content attributes used in the password policy statements ($N = 27$)	50
Table 3.4 Frequency of presentation attributes used in the password policy statements ($N = 27$)	51
Table 3.5 Frequency of language attributes used in the creation suggestion statements ($N = 11$)	52
Table 3.6 Frequency of content attributes used in the creation suggestion statements ($N = 11$)	52
Table 3.7 Frequency of presentation attributes used in the creation suggestion statements ($N = 11$)	53
Table 3.8 Frequency of media used in the password strength indicators ($N = 9$)	53
Table 3.9 Frequency of colour scheme used in the password strength indicators ($N = 9$)	54
Table 3.10 Frequency of the provision of supporting feature ($N = 30$)	57
Table 4.1 The 12 evaluated PCSs	64
Table 4.2 Overall statistics for each expert in the study	68
Table 4.3 Pairwise comparisons between experts on severity ratings of problems ...	69
Table 4.4 Cumulative percentage in levels of agreement between experts in severity ratings of problems	69
Table 4.5 Total number of usability problems found by experts per PCS with severity ratings	70
Table 4.6 Number of usability problems found by experts for each supporting feature and step	71
Table 4.7 Number of usability problems identified by experts, with the number of PCSs in which they were encountered and mean severity ratings	73
Table 5.1 The six evaluated PCSs	79

Table 5.2 Number of usability problem instances and distinct problems found by participants for each of the six PCSs, along with the mean severity ratings and range of problems per participant	83
Table 5.3 Number of distinct usability problems found by participants for each feature and step	83
Table 5.4 Number of instances of and distinct usability problems identified by users, with the number of PCSs in which they were encountered and mean severity ratings	85
Table 5.5 Frequency of the helpful best practices across six categories, with the number of PCSs in which they occurred and the number of participants who reported these practices	89
Table 5.6 Frequency of the reported worst practices across the six categories, with the number of PCSs in which they were encountered and the number of participants who reported these practices	90
Table 5.7 Frequency of the reported characteristics that should be changed across the six categories, with the number of PCSs in which they were needed and number of participants who reported these characteristics	91
Table 5.8 Number of distinct usability problems found in expert and user evaluations for each of the six PCSs, along with the mean severity ratings	92
Table 5.9 Number of distinct usability problems identified by experts only, users only, and both experts and users divided into the main and subcategories of usability problems, along with the mean severity ratings	94
Table 6.1 Demographic characteristics (frequency and %) of participants in each group and overall	102
Table 6.2 The chosen sample for the four mockup PCSs	105
Table 6.3 Mean (median) creation time, keystrokes, and perceived workload measures between the usability problems conditions	115
Table 6.4 Pairwise comparisons of creation time measure between the usability problems conditions	115
Table 6.5 Mean (median) creation time, keystrokes, and perceived workload measures between different policy presentation conditions	116
Table 6.6 Pairwise comparisons of ratings of temporal demand measure between policy presentation conditions	117
Table 6.7 Mean (median) ratings of ease of use, annoyingness, helpfulness, clarity, amount of details, and participants' confidence measures between the usability problems conditions	117
Table 6.8 Mean (median) ratings of ease of use, annoyingness, helpfulness, clarity, amount of details, and participants' confidence measures between the policy presentation conditions	118
Table 6.9 Mean (median) ratings of the password length, number of digits, number of uppercase letters, number of lowercase letters, and number of symbols measures between usability problems conditions	119

Table 6.10 Pairwise comparison for the number of digits, number of uppercase letters, and number of symbols measures between usability problems conditions	120
Table 6.11 Mean (median) ratings of the password length, number of digits, number of uppercase letters, number of lowercase letters, and number of symbols measures between policy presentation conditions	121
Table 6.12 Pairwise comparison for the number of digits, number of uppercase letters, and number of symbols measures between policy presentation conditions	122
Table 7.1 PCSs analysed for this study	137
Table 7.2 Frequency of the instructions used at the before-interaction step across the seven attributes	141
Table 7.3 Frequency of the instructions used at the during-interaction step across the seven attributes	142
Table 7.4 Frequency of the instructions used at the after-interaction step across the seven attributes	143
Table 7.5 Number of statements examined for each type of instruction across the three timings of presentation in each group	146
Table 7.6 Demographic characteristics (frequency and %) of respondents in each group and overall	148
Table 7.7 The overall structure of the questionnaire in each group	149
Table 7.8 Nature of the user instruction for each type across the three steps of interaction in each group, and the number of statement variations	152
Table 7.9 Mean (median) ratings of the dependent measures: policy at the before-interaction step	158
Table 7.10 Pairwise comparisons between the procedural policy statements at the before-interaction step across the dependent measures	158
Table 7.11 Mean (median) ratings of the dependent measures: suggestion at the before-interaction step	160
Table 7.12 Pairwise comparisons between the declarative suggestion statements at the before-interaction step across the dependent measures	161
Table 7.13 Pairwise comparisons between the procedural suggestion statements at the before-interaction step across the dependent measures	162
Table 7.14 Mean (median) ratings of the dependent measures: policy at the during-interaction step	164
Table 7.15 Pairwise comparisons between the declarative policy statements at the during-interaction step across the dependent measures	165
Table 7.16 Pairwise comparisons between the procedural policy statements at the during-interaction step across the dependent measures	166
Table 7.17 Mean (median) ratings of the dependent measures: suggestion at the during-interaction step	167
Table 7.18 Pairwise comparisons between the declarative suggestion statements at the during-interaction step across the dependent measures	167

Table 7.19 Pairwise comparisons between the procedural suggestion statements at the during-interaction step across the dependent measures	169
Table 7.20 Mean (median) ratings of the dependent measures: error message at the during-interaction step	170
Table 7.21 Pairwise comparisons between the declarative error message statements at the during-interaction step for perceived helpfulness	170
Table 7.22 Mean (median) ratings of the dependent measures: policy at the after-interaction step	171
Table 7.23 Pairwise comparisons between the declarative policy statements at the after-interaction step across the dependent measures	172
Table 7.24 Pairwise comparisons between the procedural policy statements at the after-interaction step across the dependent measures	173
Table 7.25 Mean (median) ratings of the dependent measures: suggestion at the after-interaction step	173
Table 7.26 Pairwise comparisons between the procedural suggestion statements at the after-interaction step across the dependent measures	174
Table 7.27 Mean (median) ratings of the dependent measures: error message at the after-interaction step	175
Table 7.28 Pairwise comparisons between the declarative error message statements at the after-interaction step across the dependent measures	175
Table 7.29 Summary of the current practices and user study findings	183
Table 8.1 Study design and conditions	187
Table 8.2 Demographic characteristics (frequency and %) of participants in each group and overall.....	191
Table 8.3 Password strength indicator conditions	199
Table 8.4 Mean (median) creation time and keystrokes measures between the different timing of presentation conditions for policy	203
Table 8.5 Pairwise comparisons of creation time and keystrokes measures between different timing of presentation conditions for policy	204
Table 8.6 Mean (median) ratings of ease of use, annoyingness, helpfulness, clarity, amount of detail, and participants' confidence measures between the different timings of presentation for policy	205
Table 8.7 Mean (median) Password length, number of digits, number of uppercase letters, number of lowercase letters and number of symbols measures between different timing of presentation for policy	207
Table 8.8 Pairwise comparisons of password length measures between timing of presentation conditions for policy.....	207
Table 8.9 Mean (median) creation time and keystrokes measures between the different timing of presentation conditions for suggestion.....	214
Table 8.10 Pairwise comparisons of creation time and keystrokes measures between different timing of presentation conditions for suggestion	215

Table 8.11 Mean (median) ratings of ease of use, annoyingness, helpfulness, clarity, amount of detail, and participants' confidence measures between the different timings of presentation for suggestion	216
Table 8.12 Pairwise comparisons of amount of detail and confidence measures between different timing of presentation conditions for suggestion	218
Table 8.13 Mean (median) password length, number of digits, number of uppercase letters, number of lowercase letters, and number of symbols measures between different timings of presentation for suggestion	218
Table 8.14 Pairwise comparisons of password length measures between different timing of presentation conditions for suggestion	218
Table 8.15 Mean (median) creation time and keystrokes measures for the six indicator conditions	226
Table 8.16 Pairwise comparisons of keystrokes measures between media conditions for indicator	227
Table 8.17 Mean (median) ratings of ease of use, annoyingness, helpfulness, clarity, amount of detail, and participants' confidence measures for the six indicator conditions	228
Table 8.18 Pairwise comparisons of helpfulness, clarity, and confidence measures between different media conditions for indicator	229
Table 8.19 Mean (median) password length, number of digits, number of uppercase letters, number of lowercase letters, number of symbols, and strength score measures for the six indicator conditions	232
Table 8.20 Pairwise comparisons of password length, number of digits, number of symbols, and strength score measures between media conditions for indicator	233
Table 8.21 Mean (median) recall time measure for the six indicator conditions	237
Table 8.22 Mean (median) ratings of confidence measure for the six indicator conditions	239
Table 9.1 Study design and conditions	252
Table 9.2 Demographic characteristics (frequency and %) of participants, by group and overall	255
Table 9.3 Mean (median) creation time and keystrokes measures for the four types of combination conditions	264
Table 9.4 Pairwise comparisons of creation time and keystrokes measures across the four types of combination conditions	265
Table 9.5 Mean (median) ratings of user satisfaction measures across the four types of combination conditions	265
Table 9.6 Mean (median) ratings of password characteristic measures for the four types of combination conditions	268
Table 9.7 Pairwise comparison of password characteristic measures between the four types of combination conditions	269
Table 9.8 Pairwise comparison of password strength scores across three types of combination conditions	274

Table 9.9 Mean (median) creation time measures across different presentation conditions for policy, suggestion, and indicator	279
Table 9.10 Mean (median) of keystrokes measures for policy, suggestion, and indicator across different presentation conditions	280
Table 9.11 Pairwise comparisons of keystrokes measures across different presentation conditions	280
Table 9.12 Mean (median) ratings of ease of use measures for policy, suggestion, and indicator across different presentation conditions	281
Table 9.13 Pairwise comparisons of ease of use measure ratings for policy, suggestion, and indicator across different presentation conditions	281
Table 9.14 Mean (median) ratings of annoyingness measures for policy, suggestion, and indicator across different presentation conditions	282
Table 9.15 Pairwise comparisons of annoyingness measure ratings for policy, suggestion, and indicator across different presentation conditions	282
Table 9.16 Mean (median) ratings of helpfulness measures for policy, suggestion, and indicator across different presentation conditions	283
Table 9.17 Pairwise comparisons of helpfulness measure ratings for policy, suggestion, and indicator across different presentation conditions	283
Table 9.18 Mean (median) ratings of clarity measures for policy, suggestion, and indicator across different presentation conditions	284
Table 9.19 Mean (median) ratings of amount of detail measures for policy, suggestion, and indicator across different presentation conditions	284
Table 9.20 Pairwise comparisons of amount of detail measure ratings for policy, suggestion, and indicator across different presentation conditions	285
Table 9.21 Mean (median) ratings of confidence measures for policy, suggestion, and indicator across different presentation conditions	286
Table 9.22 Pairwise comparisons of confidence measure ratings for policy across different presentation conditions	286
Table 9.23 Mean (median) of password length measures for policy, suggestion, and indicator across different presentation conditions	287
Table 9.24 Pairwise comparisons of password length measures for policy, suggestion, and indicator across different presentation conditions	287
Table 9.25 Mean (median) number of digits measures for policy, suggestion, and indicator across different presentation conditions	288
Table 9.26 Pairwise comparisons of number of digits measures for indicator across different presentation conditions	288
Table 9.27 Mean (median) number of uppercase letters measures for policy, suggestion, and indicator across different presentation conditions	289
Table 9.28 Pairwise comparisons of number of uppercase letters measures for suggestion and indicator across different presentation conditions	289

Table 9.29 Mean (median) number of lowercase letters measures for policy, suggestion, and indicator across different presentation conditions.....	290
Table 9.30 Pairwise comparisons of number of lowercase letters measures for policy across different presentation conditions.....	290
Table 9.31 Mean (median) number of symbols measures for policy, suggestion, and indicator across different presentation conditions.....	291
Table 9.32 Pairwise comparisons of number of symbols measures for indicator across different presentation conditions.....	291
Table 9.33 Pairwise comparisons of strength score measures for indicator across different presentation conditions.....	294
Table 9.34 Mean (median) of recall time measures for policy, suggestion, and indicator across different presentation conditions.....	296
Table 9.35 Pairwise comparisons of recall time measures for policy, suggestion, and indicator across different presentation conditions.....	296
Table 9.36 Mean (median) ratings of confidence measures for policy, suggestion, and indicator across different presentation conditions.....	298
Table 9.37 Pairwise comparisons of confidence measure ratings for policy, suggestion, and indicator across different presentation conditions.....	298
Table 10.1 Source and supporting data for each item in the <i>PassHeuristics</i> and <i>PassGuidelines</i>	310
Table 10.2 Components of the first heuristic and guideline and rationale for inclusion.....	312
Table 10.3 Supporting evidence from the users' perception data regarding the first heuristic and guideline.....	313
Table 10.4 Components of the second heuristic and guideline and rationale for inclusion.....	314
Table 10.5 Supporting evidence from the users' perception data regarding the second heuristic and guideline.....	315
Table 10.6 The third heuristic and guideline and rationale for inclusion.....	316
Table 10.7 Supporting evidence from the users' perception data regarding the third heuristic and guideline.....	317
Table 10.8 Components of the fourth heuristic and guideline and rationale for inclusion.....	318
Table 10.9 Supporting evidence from the users' perception data regarding the fourth heuristic and guideline.....	318
Table 10.10 Components of the fifth heuristic and guideline and rationale for inclusion.....	319
Table 10.11 Supporting evidence from the users' perception data regarding the fifth heuristic and guideline.....	320
Table 10.12 Components of the sixth heuristic and guideline and rationale for inclusion.....	321

Table 10.13 Supporting evidence from the users' perception data regarding the sixth heuristic and guideline	322
Table 10.14 Components of the seventh heuristic and guideline and rationale for inclusion.....	323
Table 10.15 Supporting evidence from the users' perception data regarding the seventh heuristic and guideline	323
Table 10.16 Second version of the <i>PassHeuristics</i> and <i>PassGuidelines</i> after the first round of feedback	327
Table 10.17 Third version of the <i>PassHeuristics</i> and <i>PassGuidelines</i> after the second round of feedback	328
Table 10.18 Mean (median) ratings of the six dependent measures for individual items of the <i>PassHeuristics</i> and <i>PassGuidelines</i>	329
Table 10.19 <i>PassHeuristics</i> to support the evaluation of PCSs.....	331
Table 10.20 <i>PassGuidelines</i> to support the development of PCSs.....	332
Table 10.21 Mean (median) ratings of the six dependent measures for individual items of the <i>PassHeuristics</i>	340

List of Figures

Figure 1.1 The three phases of this research	5
Figure 1.2 Studies conducted in (a) Phase 1, (b) Phase 2, and (c) Phase 3.....	6
Figure 1.3 Research scope	13
Figure 2.1 Relationship between the components and steps of user authentication ..	19
Figure 2.2 Examples of knowledge-based authentication mechanisms.....	20
Figure 2.3 Examples of token-based authentication mechanisms	22
Figure 2.4 Examples of biometric-based authentication mechanisms.....	23
Figure 2.5 Biometric enrolment and authentication process.....	24
Figure 2.6 Minimum to maximum number of passwords held, across seven surveys (see text) conducted between 2000 – 2012	36
Figure 3.1 Examples of the three key supporting features in PCSs.....	48
Figure 3.2 The three-step model of PCSs	56
Figure 3.3 Temporal organisation of the three key features of PCSs (% total more than 100, as features can occur at more than one step in a PCS).....	56
Figure 6.1 Structure of the password creation application used in the creation part	103
Figure 6.2 Examples of design constructs provided on the <i>Mockup1-Apple</i> password creation page across the three timings of presentation.....	108
Figure 6.3 Examples of design constructs provided on the <i>Mockup2-DailyMail</i> password creation page across the three timings of presentation	108
Figure 6.4 Examples of design constructs provided on the <i>Mockup3-Netflix</i> password creation page across the three timings of presentation.....	109
Figure 6.5 Examples of design constructs provided on the <i>Mockup4-WordPress</i> password creation page across the three timings of presentation	109
Figure 6.6 A screenshot of the <i>mockup4</i> password recall page	111
Figure 6.7 Percentage of password character classes across the four usability problems conditions	121
Figure 6.8 Percentage of password character classes across the three policy presentation conditions	123
Figure 6.9 Percentages of password guessability across the four usability problems conditions	124

Figure 6.10 Percentages of password guessability across the three the policy presentation conditions	125
Figure 6.11 Mean recall times between the four mockup PCS conditions	126
Figure 6.12 Percentages of accuracy in recalling passwords across the four mockup PCS conditions.....	127
Figure 6.13 Mean ratings of participants' confidence between the four mockup PCS conditions.....	127
Figure 7.1 The percentage of the three presentation timings for the three types of instructions.....	140
Figure 7.2 An example of the introduction page that presented the policy instruction at the before-interaction step in Group 1	149
Figure 7.3 Examples of the PCS images presented for a password policy statement across the three timings of presentation.....	150
Figure 7.4 An example of the policy statements that were investigated in the user study and how they were derived from the analysis of the instructions currently used at the before-interaction step	151
Figure 8.1 Structure of the password creation application used in the creation part	192
Figure 8.2 A screenshot of the <i>baseline</i> password creation page	193
Figure 8.3 A screenshot of the <i>policy-before-interaction</i> page	194
Figure 8.4 Screenshots of the <i>policy-during-interaction</i> page.....	195
Figure 8.5 Screenshots of the <i>policy-after-interaction</i> page.....	195
Figure 8.6 Screenshots of the <i>policy-during&after-interaction</i> page.....	196
Figure 8.7 A screenshot of the <i>suggestion-before-interaction</i> page	197
Figure 8.8 Screenshots of the <i>suggestion-during-interaction</i> page	197
Figure 8.9 Screenshots of the <i>suggestion-after-interaction</i> page	198
Figure 8.10 Screenshots of the <i>suggestion-during&after-interaction</i> page.....	198
Figure 8.11 A screenshot of the <i>3colour indicator (3colour-graphical&textual)</i> page	198
Figure 8.12 Percentage of password character classes across the four timings of presentation for policy	209
Figure 8.13 Percentage of password compliance across the four timings of presentation for policy.....	209
Figure 8.14 Percentage of password guessability across the four timings of presentation for policy	210
Figure 8.15 Mean recall times across the four timing of presentation conditions for policy.....	211
Figure 8.16 Percentages of accuracy in recalling passwords across the four timing of presentation conditions for policy.....	212
Figure 8.17 Mean ratings of participants' confidence across the four timing of presentation conditions for policy.....	212

Figure 8.18 Percentage of password character classes across the four timings of presentation for suggestion	220
Figure 8.19 Percentage of password compliance across the four timing of presentation for suggestion	221
Figure 8.20 Percentage of passwords including the given symbols across the four timings of presentation	221
Figure 8.21 Percentage of password guessability across the four timings of presentation for suggestion	222
Figure 8.22 Mean recall times across the four timing of presentation conditions for suggestion.....	223
Figure 8.23 Percentages of accuracy in recalling passwords across timing of presentation conditions for suggestion.....	223
Figure 8.24 Mean ratings of participants' confidence between four timing of presentation conditions for suggestion.....	224
Figure 8.25 Percentage of password character classes across six indicator conditions	235
Figure 8.26 Percentage of password strength levels across the six indicator conditions	236
Figure 8.27 Percentage of password guessability across the six indicator conditions	236
Figure 8.28 Percentage of accuracy in recalling passwords across the six indicator conditions	238
Figure 9.1 Structure of the password creation application used in the creation part	256
Figure 9.2 Screenshots of the design provided on the <i>policy&suggestion</i> password creation page across the three timings of presentation.....	259
Figure 9.3 Screenshots of the design provided on the <i>policy&indicator</i> password creation page across the three timings of presentation.....	259
Figure 9.4 Screenshots of the design provided on the <i>suggestion&indicator</i> password creation page across the three timings of presentation.....	260
Figure 9.5 Screenshots of the design provided on the <i>policy&suggestion&indicator</i> password creation page across the three timings of presentation	260
Figure 9.6 Percentage of password character classes across the four types of combination conditions	271
Figure 9.7 Percentage of policy compliance across three types of combination conditions	271
Figure 9.8 Percentage of suggestion compliance across three types of combination conditions	272
Figure 9.9 Percentage of passwords that included the given symbols across three types combination conditions	273
Figure 9.10 Percentage of password strength levels across three types of combination conditions	274

Figure 9.11 Percentages of password guessability across the four types of combination conditions	275
Figure 9.12 Mean recall times across the four types of combination conditions	276
Figure 9.13 Percentage of password recall accuracy across the four types of combination conditions	277
Figure 9.14 Mean ratings of participants' recall confidence across the four types of combination conditions	277
Figure 9.15 Percentage of password character classes for policy, suggestion, and indicator across different presentation conditions	292
Figure 9.16 Percentage of policy compliance across different presentation conditions	292
Figure 9.17 Percentage of suggestion compliance across different presentation conditions	293
Figure 9.18 Percentage of passwords that included the given symbols across different presentation conditions	293
Figure 9.19 Percentage of password strength levels across different presentation conditions	294
Figure 9.20 Percentage of password guessability across different presentation conditions	295
Figure 9.21 Percentage of password recall accuracy across different presentation conditions	297
Figure 10.1 Proportion (%) of covered and not covered usability problems in the <i>PassHeuristics</i>	310
Figure 10.2 Process of reviewing the versions of the <i>PassHeuristics</i> and <i>PassGuidelines</i>	324
Figure 10.3 Structure of the evaluation application	335
Figure 10.4 A screenshot of the <i>Mockup3-Netflix</i> evaluation page	336

Acknowledgments

I am most grateful to Almighty Allah for giving me the ability, determination, and strength to undertake my PhD journey. Throughout this journey, I was truly blessed for being surrounded by such wonderful people. This thesis would not be possible without them and I owe them a great deal of gratitude.

First and foremost, I would like to express my sincere gratitude to my supervisor, Professor Helen Petrie, for her incredible support, guidance, help, and valuable advice throughout my research as well as beyond my studies. I have been extremely lucky to have a supervisor who cared about her students on both personal and professional levels. I am grateful for her continuous encouragement and patience. Thank you for all the knowledge and skills you taught me during the course of my study. Thank you for being such a thoughtful and an incredible supervisor to me.

My gratitude is also extended to my examiners, Professor Karen Renaud and Dr. Christopher Power, for their time and effort in reading this thesis and for their valuable feedback. I would like also to thank them for leading a thought-provoking and enjoyable viva. I especially want to thank Dr. Christopher Power, as my internal examiner, for his valuable feedback and suggestions throughout the milestones of this research.

I would like to thank the Saudi Arabian Ministry of Education, Umm Al-Qura University, and Jeddah University for granting me a scholarship to pursue my higher education in the United Kingdom. I also would like to thank the Human-Computer Interaction research group at the University of York, led by Professor Helen Petrie for supporting my research.

I cannot forget thanking the participants for their willingness to take part in the studies. Without them, this research could not be accomplished.

There are not enough words to describe how thankful I am to my wonderful parents, amazing sisters and brother. My parents, Professor Abdullah Althubaiti and Khadijah Oqadiyah, thank you for making me who I am today and for always standing by my side to achieve my dreams. My sisters, Dr. Kholoud, Samah, Waffa, and Dr. Naseem, thank you for always pulling me up whenever I feel so down. My eternal gratitude must go to my eldest sister, Dr Kholoud, for all her help, wise advice and words of encouragement throughout this journey. My brother, Dr. Nizar, thank you for cheering me up with your sense of humour at moments I felt down and offering help whenever needed. Thank you all for your unconditional love, emotional support, and prayers. I cannot forget here my little nieces and nephews (Hashem, Numy, Mohammad, Turki, and Rahaf) who always bring laughter and happiness to my life.

Words fail me when I want to express my profound gratitude to my beloved husband, Dr. Amer Alzaidi. I am so fortunate that we both share the same field of interest in computer science. I cannot thank you enough for your expert advice on the technical aspects of my research. At the personal level, I appreciate all your support, understanding, and sacrifices all the time. You always make the best out of me and you know exactly how to make me feel comfortable not only when I need it most, but every day. Thank you for being my best friend. Thank you for being such a caring, loving, and kind husband and father to our precious son, Azzam. It is a blessing having you both in my life.

Last but not least, I would also like to thank my great parents-in-law for all their kindness, support, and nonstop prayers. I am also grateful to my friends in York and back home. Here in the UK, Fatma Layas and Sara Alotabi, I could not have wished for better friends to be by my side during this journey in making York feels like home. I also want to thank my friends back home, in particular: Ashwag Asseri and Sara Albakri for their continuous friendship and support despite long distance.

Declaration

I, *Saja Althubaiti*, declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for an award at this, or any other, University. All sources are acknowledged as References.

Some of the material contained in this thesis has appeared in the following published papers:

Althubaiti, S., and Petrie, H. (2015). *Usability Problems with Password Creation Systems: Results from Expert and User Evaluation*. Poster session presented at the Eleventh Symposium on Usable Privacy and Security (SOUPS) 2015, July 22–24, Ottawa, Canada.

Althubaiti, S. (2017). *Improving the Design and Usability of Password Creation Systems*. In Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems, CHI EA '17, pages 244–247, Denver, CO, USA. ACM.

Althubaiti, S. and Petrie, H. (2017). *Instructions for creating passwords: how do they help in password creation*. In Proceedings of 31st British Human Computer Interaction Conference, 2017, July 3-6, Sunderland, UK.

Chapter 1

Introduction

Given the rapidly growing use of digital technologies, people have increasing amounts of electronic data, which can be vulnerable to security attacks by malicious agents such as hackers, crackers, and spammers. Protecting people's data against such theft is therefore an essential precaution in the field of computer security. However, improving the security of such systems cannot be effective without taking into account the users of the systems themselves.

The field of human-computer interaction (HCI) plays an important role in addressing existing security problems. The HCI field is involved in many areas in security research and practice: user authentication, mobile security, storage, policy specification, anti-phishing efforts, device pairing, email security, and security administrators (Garfinkel & Lipford, 2014). Since the use of authentication systems is growing in line with the expansion of digital technologies, this research focuses on the authentication systems area. There are different mechanisms to authenticate users, among which the use of a personal password is the most frequent. Although passwords are widely used, however, they continue to create many problems for users and major concerns for the online security community.

Users create passwords with small interactive systems which have one or more screens including messages, strength indicators, and other elements. These password creation systems (PCSs) can be considered a particular class of interactive system that might offer supporting features to help users achieve a certain level of security during the password creation process. These supporting features primarily occur in three forms: (1) statements of password policy, which inform users about what constitutes a password; (2) password creation suggestions, which advise users on how to create

strong and memorable passwords; and (3) password strength indicators, which provide a representation of the password strength to coerced users to create stronger passwords. Although PCSs and their supporting features have existed for a long time, users still tend to choose weak passwords (Florencio & Herley, 2007).

Previous studies have focused on the strength and memorability of chosen passwords instead of looking at how supporting features are integrated into the user interface of PCSs and their effect on password choice. Evidence from Petrie and Power (2012) has shown that users encounter usability problems when creating passwords. If users struggle to understand how a PCS works, this absorbs their cognitive effort, which could otherwise be used to create a strong and yet memorable password. A recent study has shown that cognitive effort is necessary for creating passwords (Groß, Coopamootoo, & Al-Jabri, 2016).

Most users seem burdened by many textual passwords that they need to remember and use in many different systems (Grawemeyer & Johnson, 2011). They also tend to manage their passwords in insecure ways, often sacrificing security for convenience (Tam, Glassman, & Vandenwauver, 2010). To remember passwords, for example, users tend to choose an easy-to-guess one and use it with multiple accounts, or write down the strong ones because they are difficult to remember. Many studies have shown that weaknesses in passwords result from the fact that the security of passwords relies primarily on users' behaviour (e.g. Brown, Bracken, Zoccoli, & Douglas, 2004; Feldmeier & Karn, 1990; Grampp & Morris, 1984; Klein, 1990; Morris & Thompson, 1979; Sasse, Brostoff, & Weirich, 2001). Choosing a good password that is both strong and memorable, in the first instance, is the first stage of this behaviour chain.

Therefore, studying the usability of PCSs and ensuring that they support users well when creating passwords is an important issue. Considerable attention has recently been given to providing users with support for creating passwords with features within PCSs such as password strength indicators (Ur et al., 2012). However, these indicators are only one of a range of features used to encourage appropriate user behaviour in current PCSs. Apart from some preliminary work by Conlan and Tarasewich (2006), no research appears to have explored the usability of PCSs as whole interactive

systems in their own right and the usability of the support they provide to users during the creation of passwords.

1.1 Research Motivation and Aims

Considerable research attention has been paid to passwords over the past 20 years. Researchers have proposed alternative methods to replace the traditional user authentication method of passwords. Their aim has been to overcome the usability and security problems related to passwords. However, the usability of PCs as whole interactive systems on their own and the usability of the support they provide to users has been under-researched and largely ignored by researchers.

Some researchers and industry believe that passwords will not be used in the future and will be replaced by other authentication methods, such as biometric-based and token-based mechanisms (Bonneau, Herley, Oorschot, & Stajano, 2012). However, others are strongly against this belief (Herley & Van Oorschot, 2012), and the persistence of passwords gives this viewpoint credence. In 2004, Bill Gates declared that ‘the password is dead’, yet it still exists 14 years later. For now, the use of passwords is demanded each day. Billions of internet users employ passwords to access their email, social networking, and other services. According to Herley and Van Oorschot (2012), proposals to replace passwords have low expectations of success, and they are sometimes labelled as ‘yet another authentication scheme’.

Others researchers believe that password management software will be the solution to the password problem, since it provides users with complex passwords that should be stored in an encrypted database during the password creation and recall process (Stajano, Spencer, Jenkinson, & Stafford-Fraser, 2015). The password manager generates passwords for users during the creation process and then automatically fills in the passwords during the retrieval process. Although this proposal sounds promising and has been available since the 1990s, previous studies have shown the challenges that password managers bring, such as poor usability and high vulnerability (Chiasson, Van Oorschot, & Biddle, 2006; Fukumitsu, Hasegawa, Iwazaki, Sakai, & Takahashi,

2016; Garfinkel & Lipford, 2014; Li, He, Akhawa, & Song, 2014; Zhao, Yue, & Sun, 2013)

Moreover, the use of password policy enforcement tools such as nFront is increasingly being deployed as another solution to the password problem. Using such a tool forces user to comply with very strict and complex requirements to prevent the use of simple passwords. An example of such policy is to create a new password that contains a combination of different character classes, where the password should be different from previous passwords and frequently changed. However, focusing only on creating secure passwords is not a viable solution to the password problem. Creating usable passwords is as important as secure passwords. Evidence from Inglesant and Sasse (2010) indicated that enforcing such a policy will have a negative impact on users. Strict policies have been shown to affect the users' productivity. They can also make users adopt coping strategies (e.g. writing down passwords) which consequently affect the security in negative ways. Evidence from Komanduri et al. (2011) concurs with what Inglesant and Sasse (2010) have found.

Therefore, it is important to further investigate PCSs and improve the design and usability of these systems. The need to support users in choosing usable and secure passwords is clearly essential to contribute towards easing the password problem. Most of the proposed solutions have addressed the usability and security dilemma around passwords by imposing further burdens on users, such as system-generated passwords (Proctor, Lien, Vu, Schultz, & Salvendy, 2002), instead of trying to help and support them by improving the design and usability at the user interface level.

Most of what is now available in PCSs as current practices is implemented in an ad hoc manner, rather than being based on clear empirical evidence. An assessment of different PCSs revealed that there was no standard practice in employing the supporting features in such a way that helped and supported users (see Study 1, Chapter 3). One of the problems caused by this ad hoc approach is that systems have different designs of these features, which may lead to user confusion. Furthermore, current systems do not appear to provide adequate help and guidance to users in choosing a password. For example, some systems implement a password strength

indicator in their user interface but do not tell users how to increase their password strength or why the chosen password is weak.

To date, the usability of PCSs is one of the areas that has not been well studied. Exceptions are Furnell (2017) and Ur et al. (2012), who studied the role of providing a usable user interface in PCSs but only for the feature of password strength indicators. Therefore, the present research aims to inform the improvement of the design and usability of PCSs and their supporting features, in order to ensure that they support users well when creating passwords. To address this aim, the central question of this research is:

How can PCSs effectively support users in creating passwords without compromising security?

1.2 Research Approach and Methodology

The central question of this research has been addressed by breaking the research down into three phases, as shown in Figure 1.1.

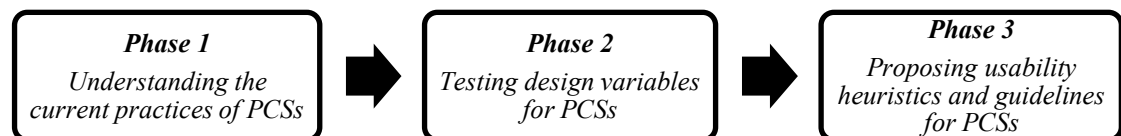


Figure 1.1 The three phases of this research

The first phase focused on understanding current practices in PCSs and their effects on users. It looked at the frequency and characteristics of the supporting features, and then investigated the usability problems they contain by conducting both expert and user evaluations of a number of PCSs. Finally, it assessed how the current practices of PCSs influence users and their passwords. The second phase investigated the effects of different design aspects in PCSs. It started by gaining a better understanding of the user instructions provided in PCSs. Then, it investigated each supporting feature individually to explore effective ways of designing the particular feature. Subsequently, it examined whether the presence of a combination of more than one supporting feature in a PCS influences the password creation process. Finally, the third

phase used the results from the previous two phases to develop and evaluate a set of usability heuristics and guidelines to support the evaluation and development of PCSs.

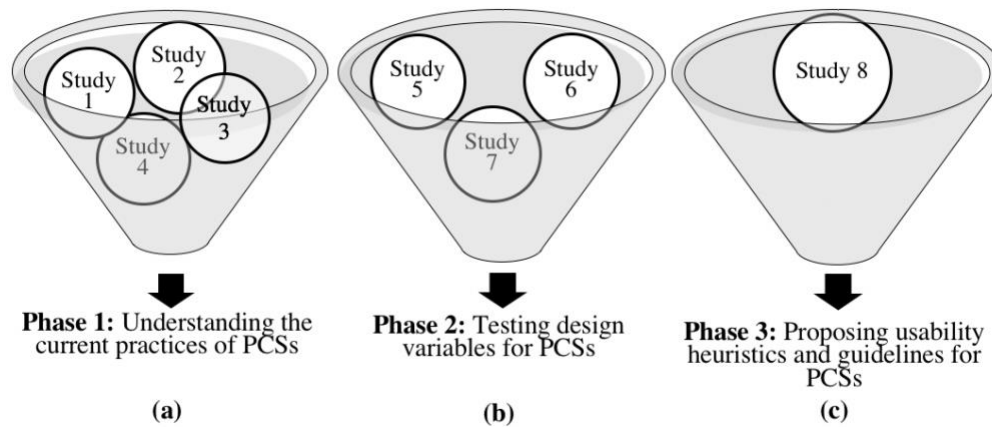


Figure 1.2 Studies conducted in (a) Phase 1, (b) Phase 2, and (c) Phase 3

The methodological approach taken in this research was based on a variety of different qualitative and quantitative methods in the area of usable security: a conceptual analysis, expert evaluation, user study in the lab, survey, and user studies online. Eight studies were conducted to address the central research question, as shown in Figure 1.2.

1.2.1 Phase 1: Understanding the Current Practices of PCSs

Four studies were conducted in Phase 1 (see Figure 1.2a): Study 1 to Study 4. The methodology used in this phase was qualitative for the first three studies and quantitative for the fourth.

The conceptual analysis method was adopted for Study 1 as it offered an effective way to understand the user interface design and its characteristics in PCSs. For Studies 2 and 3, usability testing methods were used as these are well-established methods in the field of HCI to evaluate the usability of a user interface by identifying its flaws (Lazar, Feng, & Hochheiser, 2017). However, usability testing methods have not yet been applied to the area of usable security and specifically to password research, which brings a new perspective to this topic. Finally, a controlled online experimental method was employed for Study 4. Experiments are one of the primary methods used

in a wide range of areas (Gergle & Tan, 2014). This method made it possible to see how different current practices of PCSs could influence the usability of PCSs and password strength.

The four studies are described briefly below, along with how they relate to the overall aim of the phase.

Study 1 (Chapter 3), ‘An Analysis of 30 Current PCSs’, yielded an overview of current practices in PCSs by analysing 30 PCSs. This analysis led to the conceptualisation of the password creation process into a three-step model: *before-interaction*, *during-interaction*, and *after-interaction*. The outcomes also revealed potential usability problems of current PCSs. It was therefore important to conduct usability evaluations with these systems.

Study 2 (Chapter 4), ‘An Expert Evaluation of 12 Current PCSs’, assessed the levels of usability of 12 PCSs using a collaborative expert review with seven usability experts. The experts used the three-step model as a guide through the evaluation. The number of distinct usability problems found by the experts was surprisingly high, 131 in total, even though PCSs are very small interactive systems. Therefore, it was crucial to investigate whether users would also encounter these problems.

Study 3 (Chapter 5), ‘A User Evaluation of Six Current PCSs’, assessed the levels of usability of six PCSs using a concurrent think-aloud protocol with 24 participants. A total of 654 instances of usability problems and 81 distinct usability problems were identified. Overall, most of the usability problems were related to the lack of supporting features, the timing of presentation of these supporting features (assessed in light of the three-step model), and the clarity of the instruction statements. A comparison was made between the usability problems that experts and users identified; the two evaluations produced a pool of 121 distinct usability problems: 40 (33.06%) found by experts only, 38 (31.40%) by users only, and 43 (35.54%) by both experts and users. It was therefore very important to see the impact of the usability problems and current

practices in PCSs on users and their passwords to have a comprehensive understanding of current practices of PCSs.

Study 4 (Chapter 6), ‘The Effects of Current PCS Practices on Password Creation and Recall’, examined the effects on users of current practices of PCSs using four different mock-ups with 235 participants. These mock-ups were designed based on the original designs of four current PCSs. The main finding revealed that current practices of PCSs had different effects on PCS usability and password strength. However, it was very difficult to determine a specific practice that might have caused this effect, as there was a high level of interaction between supporting features and their timing of presentation at the user interface level. Thus, the supporting features needed to be examined individually to have a clear understanding of their impact.

1.2.2 Phase 2: Testing Design Variables for PCSs

Upon completion of Phase 1, the findings suggested the need for further empirical investigation in terms of the effects of the design of supporting features. Therefore, three more studies (Study 5 to 7) were conducted (see Figure 1.2b). The methodology used in this phase was a mainly quantitative approach for all three studies.

Content analysis and survey methods were adopted for Study 5. The latter method provided a reliable measure of people’s attitude and experience (Müller, Sedley, & Ferrall-Nunge, 2014), specifically in this study, regarding the instruction statements used in current PCSs. On the other hand, a controlled online experiment method was employed for both Studies 6 and 7. Using a controlled experiment helped to examine different design variables for individual (Study 6) and combined supporting features (Study 7) and to investigate how they could influence the usability of PCSs and password strength.

The three studies are described briefly below, along with how they relate to the overall aim of the phase of research.

Study 5 (Chapter 7), ‘Instructions for Creating Passwords: Analysis and User Study’, examined what forms of instructions users prefer for the statements of password policy, creation suggestions, and error messages. The study first analysed a total of 95 existing instructions to support users in creating passwords from 27 current PCSs. Then, an online questionnaire study was conducted based on this analysis with 117 respondents to understand how the most frequently used instructions affect users’ perceptions of the instructions. The main finding was that current practices of user instructions vary widely and do not match users’ needs.

Study 6 (Chapter 8), ‘The Individual Effects of Supporting Features on Password Creation and Recall’, examined each supporting feature (policy, creation suggestions, and strength indicator) individually in a PCS by manipulating their presentations in a controlled online experimental study with 257 participants. It specifically looked at the effects of different timing of presentations (in light of the three-step model) for password policy and creation suggestion. It also investigated the effects of different media and colour scheme presentations for the strength indicator. In general, the findings suggested that different presentations of the supporting features affected the PCS usability and password strength differently when users created passwords, but not when they recalled them. Another finding showed that the mere presence of supporting features affected the PCS usability and password strength during the creation and recall processes.

Study 7 (Chapter 9), ‘The Combined Effects of Supporting Features on Password Creation and Recall’, investigated the effects of presenting more than one supporting feature simultaneously in a PCS in a controlled online experiment with 220 participants. The study used the outcomes identified in Study 6 for each supporting feature to design four combinations of supporting features. In general, the findings were similar to those found in Study 6. Different combinations of supporting features affected the PCS usability and the password strength when users created passwords, but not when they recalled them. In addition, the mere

presence of combined supporting features affected the PCS usability and password strength only when users created passwords. Finally, the individual (Study 6) and combined (Study 7) presentations of supporting features were compared, which revealed that having more than one supporting feature in the PCS improved the user's satisfaction.

1.2.3 Phase 3: Proposing Usability Heuristics and Guidelines for PCSs

Based on the findings of the first two phases, the last study (Study 8) proposed a set of heuristics for evaluators and guidelines for developers (see Figure 1.2c) specifically for the evaluation and development of PCSs. The methodology used in this phase included both quantitative and qualitative approaches. A content analysis method was used for the development process, while an online expert review was used for the evaluation process of the heuristics.

Study 8 (Chapter 10), 'Password Creation System Heuristics and Guidelines: Development and Evaluation', aimed to develop and evaluate a set of usability heuristics and guidelines to support the evaluation and development of PCSs. These guidelines and heuristics were grounded in empirical data, specifically from the usability problems users experienced and their perceptions of current PCSs (Study 3), in addition to supporting evidence from the experimental data (Studies 5, 6, and 7). After three rounds of thorough feedback, a set of 10 heuristics and guidelines for PCSs were defined, hereinafter called *PassHeuristics* and *PassGuidelines*, respectfully. The results of the evaluation of *PassHeuristics* showed that the nine evaluators identified an average of 12.22 usability problems for each mock-up. Furthermore, the evaluators intended to use *PassHeuristics* in the future, as it covered all aspects of PCSs and was perceived as easy to use and useful.

1.3 Research Validity

To ensure the validity of the research findings, a number of considerations were taken into account regarding internal and external validity. A mixture of qualitative and quantitative approaches was used throughout this research. For the qualitative approach, a Cohen's Kappa measure was used to assess the subjective interpretation of the content analysis. For the quantitative approach, open-ended questions were included to give participants the chance to explain or express their thoughts about the studies, and at the same time to provide better insight into the results.

A large number of PCSs were selected (30) from the Alexa top 100 most visited websites for analysis to ensure that the chosen sample was representative of current PCSs. These PCSs varied across a wide range of domains. While conducting Phase 1, the websites that provided the PCSs were checked regularly to see whether there were any dramatic changes in the PCSs that needed to be taken into account; none of the chosen PCSs changed during that time. Once Phase 2 started, an updated list of PCSs (specifically for Study 5) was collected from the Alexa top 100 most visited websites.

All studies were piloted to test their overall design and procedure. The clarity of task instructions and whether they were followed correctly was one of the measurements during the pilot. For the online experiments specifically, the pilot study was also used to check for technical problems. In most of the studies, some concerns were raised about either the procedure or the materials. Consequently, adjustments were made to address these concerns. The data from the pilot sessions in all studies (except Study 2) were not included in the data analysis, and participants were not allowed to take part in the main studies.

All controlled experiments and surveys were conducted online using the crowdsourcing platform Amazon Mechanical Turk (MTurk). The use of online research provided access to a large pool of participants and yielded a high response rate. Participants were not allowed to take part in more than one study to avoid any practice effects. As there was no direct interaction between participant and researcher in the online studies, this should mean that participant bias in terms of creating strong

passwords as a way to please the researcher (Garfinkel & Lipford, 2014) should have been low. The quality of the data collected from MTurk was tested in Study 5, by collecting half the data through MTurk and half through other recruitment methods. The results showed no significant differences between the two methods. Another measure was also applied to check the quality of the data in MTurk: participants were asked to answer open-ended questions to check all their answers to the questions were relevant and informative, instead of only completing a task and giving their ratings.

For ethical considerations of security and privacy, all participants were asked to create fictitious passwords and not to use their own passwords during the password creation process. However, to improve the ecological validity of the task, a scenario-based approach was used with an online bank account context. A recall task was also included in the studies' design to make participants more vigilant about their newly created passwords.

1.4 Research Scope

The main aim of this research is to inform the design and usability of the PCSs and their supporting features, and to ensure that they support users well when creating passwords. This thesis addresses this specific goal at the user interface level by providing knowledge about how users react to a range of aspects of the supporting features in PCSs and by providing a set of usability heuristics and guidelines that support the evaluation and development of existing PCSs. It is not part of this research to propose alternative designs to replace current PCSs.

Figure 1.3 illustrates the scope of the research. As discussed earlier, one of the major themes of research in the area of usable security is user authentication (Garfinkel & Lipford, 2014). Traditional passwords remain the most common mechanism to authenticate users in a system. A general review of the existing user authentication mechanisms is covered in the literature review (see Chapter 2) along with a detailed review of research on textual passwords. Of all the different types of passwords, this research investigates textual passwords that consist of characters only, such as letters,

numbers, and symbols. Throughout the thesis, these are referred to as passwords for short.

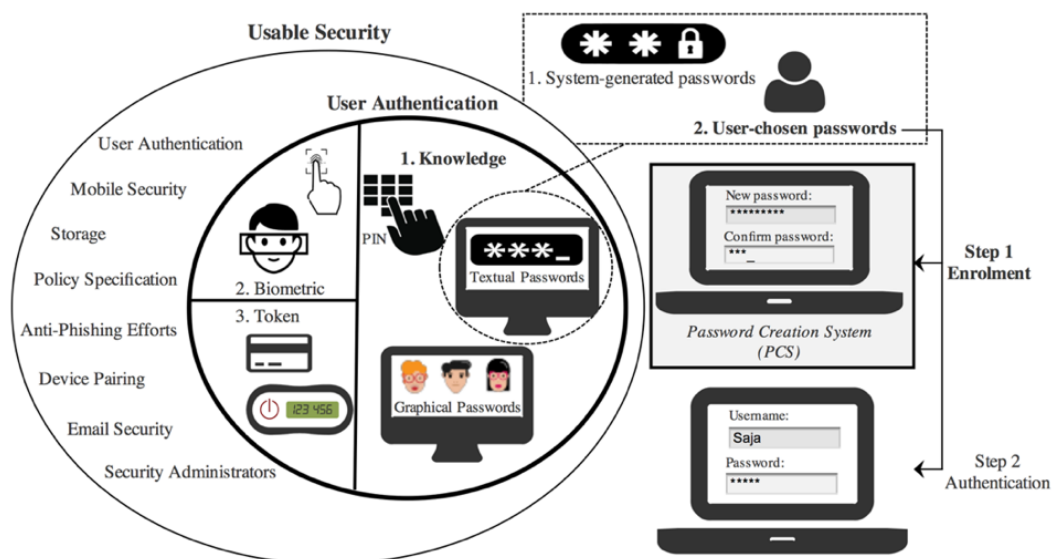


Figure 1.3 Research scope

As this research began by understanding the practices of existing PCSs, it is essential to identify the scope of the PCSs under investigation. System providers have to choose between two schemes when they decide to use a password as a medium of authentication for their website: system-generated or user-chosen passwords. In the first scheme, users are given a random password by the system and have to remember it. However, the present thesis is concerned with the second scheme, in which users are given the freedom to choose their own passwords and consequently remember them.

In general, users interact with passwords for password-protected websites at three different stages: enrolment (password creation process), authentication (password recalling process), and recovery (password resetting process). Users' interaction with PCSs at the enrolment stage occurs when they visit the website for the first time and want to register by creating a new password; the authentication stage takes place frequently, every time they want to return to the website and log in; and finally, the recovery stage occurs when they forget their password and need to create a new one.

Although both the enrolment and recovery stages involve creating a password (technically a new password for the enrolment stage and another one for the recovery stage), the present researcher found that websites often deploy different systems and present different supporting features in these stages. Therefore, the decision was made to study only the enrolment stage, hereafter called the password creation process, as it is when the user first makes their password for the system.

1.5 Research Contributions

This research makes several contributions to password research in the field of usable security, and specifically PCS usability. To the best of the author's knowledge, this is the first research on the usability of PCSs as a whole interactive system in their own right and the usability of the support that they provide to users in the creation of passwords. The main contributions of this research are summarised below:

- The research provides an understanding of the problems that people encounter when creating passwords by collecting a corpus of PCS usability problems through user and expert evaluations (Studies 2 and 3 in Chapters 4 and 5, respectively).
- The research provides an understanding of user instructions in the field of password research (Study 5 in Chapter 7).
- The research examines how PCSs should design and implement their supporting features to improve usability and password strength through user studies that manipulate these features (Studies 6 and 7 in Chapters 8 and 9, respectively).
- The research provides a set of usability heuristics and guidelines for use in guiding the evaluation and development of PCSs (Study 8 in Chapter 10).

In addition, this research makes the following secondary contributions:

- The research uses a variety of usability evaluation methods for the first time in the field of usable security (Studies 2, 3, and 8 in Chapters 4, 5, and 10, respectively).

- The research confirms previous findings and provides additional evidence that the presence of supporting features affects PCS usability and password strength (Studies 6 and 7 in Chapters 8 and 9, respectively).
- The research provides additional evidence regarding users' common practices in password creation and recall (Studies 4, 5, 6, and 7 in Chapters 6, 7, 8, and 9, respectively).

1.6 Statement of Ethics

All studies involving participants in this research were ethically approved by the Physical Science Ethics Committee of the University of York. In addition, they were designed based on the ethical considerations: do no harm, informed consent, and confidentiality of data.

Do no harm. None of the participants were put in any harmful or risky situations in any of the studies. Participants were asked to create a new password that they had never used before to avoid compromising their security. Moreover, they were told that the created passwords would not be used in any real systems.

Informed consent. Participants were informed about the study, the tasks involved, and the duration of the study during the recruiting process. The informed consent statement/form involved the following information: the researcher's and supervisor's names, the purpose of the study, tasks involved, how long the study would take, confidentiality of the data, the right to withdraw from the study, and contact information. For the online studies, all participants began their application by completing an informed consent statement (see Appendix A, Section A.1) whereas in the lab studies, all participants signed an informed consent form (see Appendix A, Section A.2).

Confidentiality of data. All data gathered in all studies were completely anonymised. Only the researcher and her supervisor have seen the data. The information collected has not be assigned to any particular participant. Since the raw data of all studies consist of passwords, this thesis does not describe any formation patterns that

participants used to create their passwords. In other words, any information that could help an attacker to easily break passwords is not provided in this work.

1.7 Thesis Structure

The rest of this thesis is organised as follows. Chapter 2 outlines the literature review. Subsequently, Chapter 3 presents an analysis of 30 current PCSs (Study 1), Chapter 4 reports an expert evaluation of 12 current PCSs (Study 2), and Chapter 5 describes a user evaluation of six current PCSs (Study 3). Chapter 6 then examines the effects of current PCS practices on password creation and recall (Study 4), while Chapter 7 presents an analysis and user study regarding instructions for creating passwords (Study 5). Next, Chapter 8 describes the individual effects of supporting features on password creation and recall (Study 6), and Chapter 9 examines the combined effects of these features on those processes (Study 7). Chapter 10 then presents the development and validation of a set of password creation system guidelines and heuristics (Study 8). Finally, Chapter 11 concludes the thesis and makes recommendations for future work in this field.

Chapter 2

Literature Review

2.1 Introduction

The aim of this chapter is to review and evaluate previous work on authentication systems. First, the chapter discusses existing user authentication mechanisms in general. It then provides a detailed review of the most common type of user authentication, textual passwords. Subsequently, the chapter covers the challenges and coping strategies users face related to passwords, and finally provides an overview of password strength and security threats.

2.2 User Authentication

The purpose of this section is to review the literature on user authentication. It begins by defining and classifying existing user authentication mechanisms. Then, it defines and discusses the most common authentication mechanisms—knowledge-, token-, and biometric-based authentication—in terms of their procedures, strengths, and weaknesses.

2.2.1 Definition of User Authentication

In the field of computer security, user authentication is defined as a mechanism used to verify the identity of an individual making a request to access a system (Renaud, 2005). However, user authentication is not a new concept invented for the world of computers: it dates back at least to the 18th century. According to Smith (2001), it may have first appeared in the folk tale ‘Ali Baba and the Forty Thieves’. In this story, Ali Baba needs to say the magic words ‘Open Sesame’, a spoken password, to gain access to treasure hidden in a cave behind a stone. Regardless of whether passwords

to gain access are computer-based or not, it is essential to understand the components and steps of a user authentication system. Smith (2001) lists five components common to most authentication systems:

- 1) a *person*, who needs to be authenticated;
- 2) a *distinguishing characteristic*, which helps to distinguish that particular person from others;
- 3) an *authentication mechanism*, which is used to verify whether the distinguishing characteristic belongs to the person attempting to use it;
- 4) an *access control mechanism*, which is responsible for granting privileges in the case of successful authentication; and finally
- 5) a *proprietor*, who is in charge of the authentication system and determines what the distinguishing characteristic should be and what mechanism to use for authentication.

In contrast to Smith (2001), Renaud (2005) proposes a three distinct steps of user authentication based solely on processes in computer-based systems: *identification*, *authentication*, and *authorisation*. Identification is a step in which a system identifies who has claimed a particular identity, usually by means of an account username. Second, the authentication step verifies whether the claimed identity is in fact the owner of that particular account. A broad range of authentication mechanisms can be used to check claimed identities, such as passwords, key fobs, and fingerprints. Finally, the authorisation step allows an authenticated user to access a system and grants a set of privileges based on his or her identity.

Based on Smith's (2001) and Renaud's (2005) concepts in terms of the components and steps involved in user authentication, the present author created a comprehensive model of user authentication to show the relationship between the components and steps in user authentication, as illustrated in Figure 2.1. As shown in the figure, the components *person* and *distinguishing characteristic* form the core of the first step in user authentication, which is *identification*. As can be expected, the *authentication mechanism* is the main component of the *authentication* step. Finally, the *access control mechanism* component plays a vital role in the *authorisation* step. The

proprietor component is not part of a particular step, but instead represents the individual or organisation responsible for the whole process.

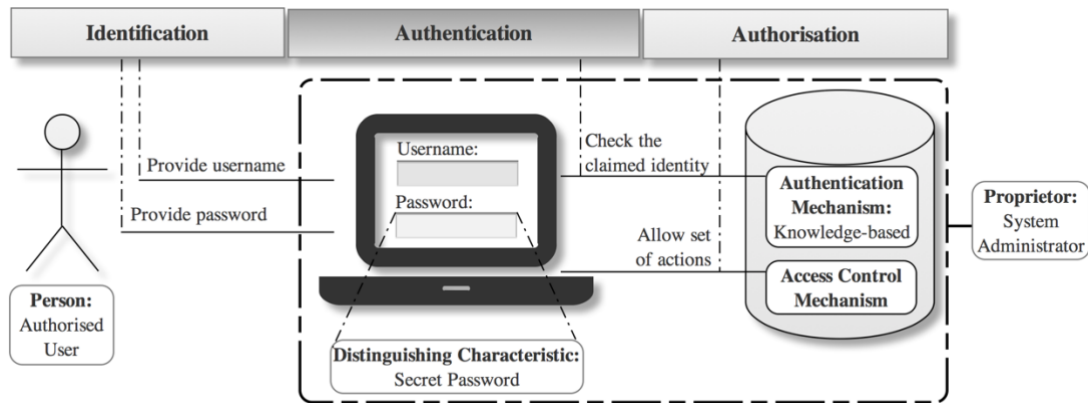


Figure 2.1 Relationship between the components and steps of user authentication

Three different stages should be considered in user authentication. The first stage in each authentication mechanism is known as the *enrolment/creation* stage. In this stage, the users must choose and set up their distinguishing characteristic (e.g. password) that will be used in the second stage. The second stage is *authentication*, as presented by Smith (2001) and Renaud (2005). In the third stage, the users can change their distinguishing characteristic (e.g. password) if needed.

In the literature, *authentication mechanisms* are usually classified in terms of factors (also known as unique characteristics) that are employed to authenticate users. Carlton, Taylor, and Wyszynski (1988) classify these factors into three main types: (1) ‘something the users know’ (knowledge-based authentication), (2) ‘something the users have’ (token-based authentication), and (3) ‘something the users are’ (biometric-based authentication). Since they were first proposed in the 1980s, these three factors have continued to be the most commonly used ways to ascertain user uniqueness. In reality, the token-based authentication is almost combined with knowledge-based authentication, in which the token claims the identity whereas the knowledge authenticates the claimed identity. Nevertheless, some researchers have proposed new factors to authenticate users, such as ‘where is the user located’ (location-based authentication) (Denning & MacDoran, 1996), ‘what is the user’s motion’ (movement-

based authentication) (Chong & Marsden, 2009), and ‘who the users know’ (Brainard, Juels, Rivest, Szydlo, & Yung, 2006).

The following sections address each of the three main types of user authentication mechanisms in detail.

2.2.2 Knowledge-Based Authentication Mechanism

The term knowledge-based authentication tends to be used to refer to any authentication mechanism that relies on users providing some knowledge¹ that they hopefully keep secret. A considerable amount of work has been published on various types of secret knowledge.

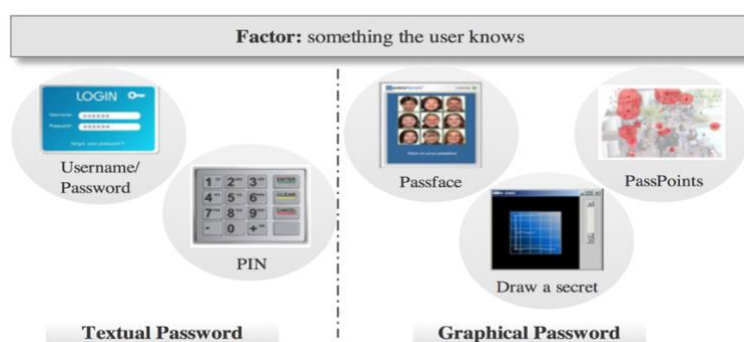


Figure 2.2 Examples of knowledge-based authentication mechanisms

As shown in Figure 2.2, textual and graphical passwords are well-known examples of secret knowledge. This secret knowledge is shared only between a particular user and the authentication system, and both sides should thereafter keep that knowledge safe for security reasons. During the *enrolment* step, the knowledge-based authentication system may either assign a password to users (i.e. system-assigned) or require users

In fact, information would be a more appropriate term than knowledge in this context, as it refers to information the user has, not his or her knowledge of a subject. However, as knowledge-based authentication is now a well-established term, the present author will continue to use the term knowledge in this context. ¹

to choose their own (i.e. user-chosen). In both cases, the user needs to memorise the password and confirm it to authenticate to the system.

Despite the wide range of options for authentication systems outlined above, knowledge-based authentication remains the most common mechanism (Herley, van Oorschot, & Patrick, 2009). There are several reasons for this, as highlighted by Carlton et al. (1988) and Smith (2001). From the system point of view, it is relatively easy and inexpensive to implement. It also does not require any additional hardware. Furthermore, knowledge-based authentication benefits not only the system side but also the user side. Most users are now familiar with the working of this kind of authentication, and they can authenticate themselves without worrying about privacy issues, unlike with biometric-based authentication (Biddle, Chiasson, & Van Oorschot, 2012). Knowledge-based authentication can also be highly suitable for users who roam regularly and need to log in to a system from different locations, unlike token-based authentication, where users have to carry physical devices. On the other hand, users have difficulty in remembering both system-assigned and user-chosen passwords (e.g. Jeff Yan, Blackwell, Anderson, & Grant, 2004). As a result, there is an increase in the cost of resetting forgotten passwords (e.g. by calling the help desk). To address this problem, users tend to use coping strategies such as writing down their passwords or choosing passwords that are easy to guess (e.g. Adams & Sasse, 1999; Brown et al., 2004; Dhamija & Perrig, 2000; Florencio & Herley, 2007; Grawemeyer & Johnson, 2011). According to Smith (2001), writing down the password negatively affects security: easy-to-guess passwords can be hacked relatively easily. In other words, there are serious usability and security deficiencies associated with knowledge-based authentication. These are covered in more detail in Section 2.3.

2.2.3 Token-Based Authentication Mechanism

The term token-based authentication refers to the situation in which users have to use a physical object, or a token, to identify themselves. Examples of such tokens are key fobs, infrared card readers, and smartcards, as shown in Figure 2.3. During *enrolment*, the system administrator provides a token to legitimate users. Users then need to

present and submit this token to the system for identification using an appropriate process.

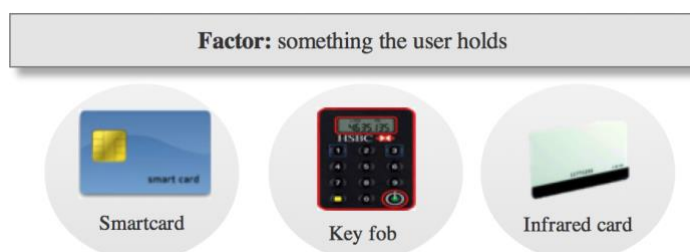


Figure 2.3 Examples of token-based authentication mechanisms

The strengths and weaknesses of using token-based authentication are as follows. The main strength of this mechanism is the low demand on users' memory. Unlike in knowledge-based authentication, users do not need to remember different passwords for each system when they log in; they must only remember to carry their tokens with them. Furthermore, concerning security, tokens may be the most difficult to abuse (R. E. Smith, 2001) for several reasons. First, this mechanism relies on one user owning a unique token. Second, a compromised token can be easily detected. In other words, users can tell when their token has been lost or stolen, and consequently the system can deactivate it. Carlton et al. (1988) emphasise that the use of tokens relies on protecting the token: the systems verify only the validity of the presented tokens, not who the holder is (Tan, Hsu, & Pinn, 2001). Moreover, the portability of tokens is both a strength and a weakness. For example, tokens can be shared between valid users in some circumstances (e.g. users who need assistance), but they can also be targets for theft as those living close to a person can easily use the tokens without the person knowing that. In addition, tokens could cause inconvenience to users; for example, roaming users need to save space in their purse or pocket to carry their token. From the system point of view, implementing tokens incurs higher costs. The effects of missing tokens are similar to those of forgotten passwords: in both cases, the users must seek help, for example by contacting the help desk.

Because tokens can be stolen or lost, token-based authentication is commonly combined with other authentication mechanisms (such as knowledge-based

authentication). This strategy is called two-factor authentication, in which the user is asked to perform multiple types of authentication. For example, automated teller machines (ATMs) typically use such a mechanism. For identification, the ATM asks the user to insert a personal bank card as a token, and the user is then asked to enter the correct PIN for verification. Two-factor authentication increases security, but it also affects usability because users have to remember both tokens and passwords.

2.2.4 Biometric-Based Authentication Mechanism

Biometric-based authentication is a mechanism that relies on measuring some intrinsic feature or features of the user. As shown in Figure 2.4, the features can be either physiological or behavioural. Physiological features include fingerprint, eye iris, face, and hand, whereas behavioural features include signature, voice, and keystroke (B. Miller, 1994).

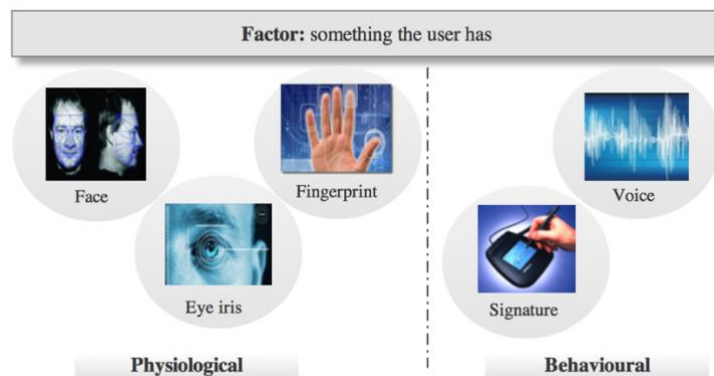


Figure 2.4 Examples of biometric-based authentication mechanisms

During *enrolment*, users must register with the system for their biometric data to be captured. Once the biometric data have been collected, a digital template of those data is created and stored in a database. During authentication, the users present their biometric data for identification. The system then verifies their identity using pattern recognition by capturing their biometrics, extracting the features from the biometric data, and comparing the features with the digital template in the database (Jain, Ross, & Prabhakar, 2004). Figure 2.5 shows the four components of a biometric system and how they relate to one another: sensor, feature extraction, matcher, and database.

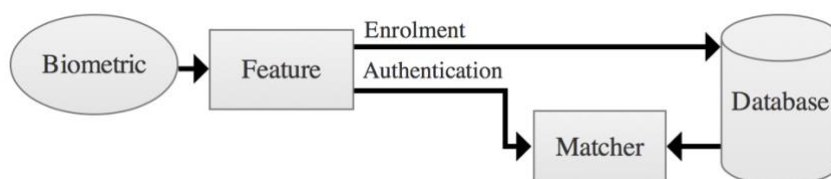


Figure 2.5 Biometric enrolment and authentication process

As long as the biometric features are distinctive for each person, biometric-based authentication is more capable of distinguishing between authorised and unauthorised people than knowledge- or token-based authentication, and more reliable in doing so (Jain, Hong, & Pankanti, 2000). There is barely any possibility of passing or sharing the means of authentication with other users in biometric-based authentication. In addition, users do not have to remember their biometric features. On the basis of the explanation above, biometric applications can be used for both identification and verification processes. During identification, the captured biometrics are checked against all the templates in the database, whereas in the verification process, the captured biometrics are checked against a specific template (Coventry, 2005).

However, there are also disadvantages to using biometric-based authentication. For example, physiological biometric characteristics can be stolen. An attacker can gain access to the system by copying an authorised user's characteristics from the fingerprint reader. In addition, biometrics are not always available or immutable. Women lose their fingerprints when they age. People who work outdoors and damage their hands often do not have viable fingerprints. There are also some privacy concerns related to biometrics such as the safety and misuse of the data. The possibility of false acceptance or false rejection is another disadvantage. Injuries, for instance, can affect users' biometric readings and lead consequently to false rejection for authorised users. From the system point of view, implementing a biometric-based system is expensive, since it requires special hardware. In addition, biometric characteristics cannot be updated, unlike passwords, which can be reset (Schneier, 1999).

2.3 Knowledge-Based Authentication: Textual Passwords

This section reviews textual passwords in terms of their types and current problems.

2.3.1 Definition of Textual Passwords

The term *textual password* refers to any secret knowledge used in identification and authentication that consists of a sequence of characters. This is also known as a character-based password. According to Renaud (2005), knowledge-based authentication mechanisms may implement one of two textual password approaches: *random* or *cultural*. The random password approach tends to be used more often, and consists of a sequence of characters. In this approach, the system assigns passwords to users or requires users to choose their own passwords. This is the approach used in most computer-based environments. However, depending on the type of random approach, the textual password is given different names. For instance, it is called a *password* when it consists of characters only (such as letters, numbers, and symbols), a *personal identification number* (PIN) when it consists of digits only, and a *passphrase* when it consists of a phrase.

In contrast, the cultural password approach tends to rely on a rational process to create a required password. During the *enrolment* step, the system typically asks the user to answer one or more challenging questions (e.g. ‘Which primary school did you attend?’). The answer to these challenging questions is related to an established fact, interest, or opinion of the user. Therefore, passwords based on a cultural approach are also known as *cognitive* or *semantic* passwords. However, most computer-based systems only use the cultural approach to recover forgotten passwords, as noted by Just (2005). An example of this approach is personal verification questions (PVQs).

2.3.2 Problems with Textual Passwords

Although the textual password is the primary means used for authentication, it has a number of usability and security problems (Jianxin Yan, Blackwell, Anderson, & Grant, 2000). Some textual passwords, while easy to remember, are also easy to guess. In contrast, other textual passwords are highly difficult to guess but also to remember. According to Wiedenbeck, Waters, Birget, Brodskiy, and Memon (2005), the *password problem* refers to the situation where the password can be either memorable but insecure or secure but difficult to remember. Furthermore, the authors claim that

it is impossible to fulfil the requirements of memorability and security for users despite the need for both. Users also tend to manage their passwords in insecure ways, and often have to sacrifice security for convenience (Tam et al., 2010). To remember passwords, for example, users tend to choose an easy-to-guess password and use it with multiple accounts, or write down the strong² ones. Many studies have shown that most password weaknesses result from the fact that password security relies primarily on users' behaviours (e.g. Brown et al., 2004; Feldmeier & Karn, 1990; Klein, 1990; Morris & Thompson, 1979; Sasse et al., 2001).

Users' behaviours that are linked to the password problem may be seen from two perspectives:

1. The system side: examining when users create their passwords and how they do so by examining their password choice and the systems they use for password creation (see Section 2.4); or
2. The user side: examining how users handle their passwords by understanding their password-related behaviours (see Section 2.5).

Of all the different types of passwords, this research investigates textual passwords that consist of characters only. Throughout the thesis, it is referred to as passwords for short.

2.4 Password Creation Systems

Users create passwords with small interactive systems consisting of one or more screens, which might include messages, strength indicators, and other elements. Such password creation systems (PCSs) are considered as a particular class of interactive system that may offer supporting features to help users achieve a certain level of security during the password creation process. These supporting features include

² A 'strong' password is defined as one that is not easy to guess.

statements of password policy, statements of password creation suggestions, and password strength indicators.

2.4.1 Password Choice

In the field of computer security, system security attracted significant attention after national and international networks were at risk of being breached in the late 1980s. Several important questions arose when a worm program³ cracked users' passwords using a mini-dictionary with only 432 words (Spafford, 1989). One of these questions was whether the list of 432 words was representative of passwords on other computers (Riddle, Miron, & Semo, 1989). Thus, user-chosen passwords have been analysed in terms of their intrinsic attributes over the past 20 years.

The findings of these analyses generally show that users are likely to choose passwords on the basis of familiar words to remember them (Riddle et al., 1989). Furthermore, passwords that can be derived from a dictionary are more vulnerable to guessing (De Alvare, Schultz, & Ne, 1988). This information, while well known, is essential in the field of computer security.

Researchers have studied the problem of choosing good passwords and how to make passwords more difficult to guess. In a classic study by Grampp and Morris (1984) on UNIX security, users were forced to choose new passwords for new system. The new passwords should contain at least six characters and at least one digit; the system rejected any passwords consisting purely of letters or digits. The authors then examined several dozen systems using trial passwords. These trial passwords were a collection of 20 common female names, each followed by a single digit. Out of 200 passwords, at least one was in use on each examined machine. However, it might be

³ A *worm* is a program that can run and replicate itself in other machines. It is different from a *virus*, which is a code that adds itself to other programs.

that users in this study simply reused one of their passwords or used a password manager to do the job for them.

Klein (1990) conducted a well-known study on password vulnerability to illustrate which password attributes make a password more vulnerable to guessing attacks. He collected a database of 15,000 UNIX account entries in encrypted files. These entries belonged to his friends and acquaintances. Each account entry was tested using different methods of attack to see whether the passwords used were vulnerable to compromise. Most of the attack methods were based on a user's name or account number. The results showed that a quarter of the passwords could be cracked. Furthermore, depending on the amount of effort put into choosing the passwords, Klein expected that by the end of the first day, between 5 and 15 accounts would be cracked on an average system with 50 accounts. In his study, the most probable choices of passwords were those taken from dictionary words (7.4% of passwords), common names (4.0% of passwords), and combinations of the user and account names (2.7% of passwords).

More recently, Shay et al. (2010) conducted a study of 470 computer users who had changed their passwords to comply with new policy requirements. Their results confirmed that dictionary words (42.7%) and names (34.6%) were still the most common strategies for creating passwords. In contrast, the less probable choices used as a basis for passwords are public information (11.9%) and mnemonic style (5.5%). In addition, they found that less than 30.0% of users composed an entirely new password each time, with 52.4% of users modifying an old one instead.

In a diary study, Grawemeyer and Johnson (2011) asked participants to keep a note of the characteristics of their passwords. Their findings identified seven types of created passwords: (1) a single/common word or name, (2) a meaningful phrase, (3) an abbreviation of a meaningful phrase, (4) a meaningful combination of letters and numbers, (5) a number pattern, (6) random characters, or (7) another pattern. Out of the 991 passwords, most contained an abbreviation of a meaningful phrase (27.7%), followed by a single/common word or name (25.9%). In contrast, few passwords contained random characters (12.7%), other patterns (12.0%), a meaningful

combination of letters and numbers (9.0%), a meaningful phrase (7.2%), or a number pattern (5.4%). However, according to Yan et al. (2004), passwords based on either meaningful phrases or random characters are the most secure.

The US Department of Defense (1985) highly-recommends random system-assigned passwords over user-chosen passwords because the former are much more complex. On the other hand, the National Institute of Standards and Technology (NIST) (2017) recommends the use of user-chosen passwords in the form of passphrases. In the case of user-chosen passwords, some users tend to put little effort into choosing password content. Research by Florencio and Herley (2007) provides supporting evidence that users opt for simple passwords over complex ones. For instance, assessing the passwords of half a million users, they found that the majority of users choose passwords that contain lowercase letters only.

2.4.2 Supporting Features for Creating Passwords

Three supporting features can be implemented in PCSs to help users create strong and yet memorable passwords. The first supporting feature is the password policy, which is a set of constraints that determine the content and general use of all passwords for a given system. The second feature is password creation suggestions, which advise users on the content of good and secure passwords and/or password structure without enforcing users to comply with it. Finally, the third supporting feature is the password strength indicator, which is a visualisation of an estimate of the strength of a proposed password.

2.4.2.1 Password Policy

Morris and Thompson (1979) believed that giving free choice to users in creating their passwords would make an attacker's job much easier. Therefore, they conducted a study to determine user password choosing habits. Their results showed that users chose passwords from a restricted set of characters, such as all lowercase letters or all

digits. As a result, these authors were the first to propose the idea of a system that enforced a policy for choosing passwords.

Six years later, the Federal Information Processing Standards (FIPS) (1985) outlined such a standard on passwords usage in terms of the content for users and system administrators. In addition, the FIPS suggested password system requirements for different levels of security that afford low, medium, and high protection. According to the FIPS, the following 10 factors should be considered when designing and implementing a password system:

1. *Composition* (or character classes): the acceptable types of characters that form the password, such as lowercase letters, uppercase letters, digits, and symbols.
2. *Length* range: a determination of the range of the acceptable number of characters.
3. *Lifetime*: the acceptable time period during which a password may be used.
4. *Source*: the entities that can create or select passwords, either the owner or a password generator.
5. *Ownership*: the individual who owns and uses the password.
6. *Distribution*: the method used for distributing a new password from the owner to the place in the system.
7. *Storage*: the method used to store used passwords.
8. *Entry*: the method with which a password may be entered by an automated data processing user.
9. *Transmission*: the method used for communicating a password once it is entered for comparison with a stored password.
10. *Authentication* period: the maximum period of time between any initial authentication process and the re-authentication process during sessions.

Typically, these factors are addressed by the system administrator to encourage users to generate passwords that are more difficult to crack. However, some clarification is necessary about what users perceive as relevant in policies when choosing passwords. Two factors among these 10 relate to the stage of choosing a password: *composition* and *length*. Regarding security, a PCS sometimes provides a password policy on-

screen to users during the password creation process. Furthermore, PCSs implement password policies by using proactive password checking, so passwords are not accepted unless they satisfy the existing policy.

However, according to Sasse et al. (2001), password policy increased the password problem, since most of the password policies used are based on the FIPS guidelines proposed in 1985. Their results show that the major cause of password problems is forced password changes.

Proctor, Lien, Vu, Schultz, and Salvendy (2002) examined the effectiveness of enforcing a password policy on users and their passwords. Their results revealed that this improved the strength of the passwords and did not affect the accuracy of the recalled passwords. However, it also increased the user's difficulty in creating an acceptable password, occurring mainly in the time needed to create a password that satisfied the policy. Furthermore, the impact of providing a password policy on improving the strength of the passwords has been proven in the literature (e.g. Campbell, Ma, & Kleeman, 2011; Vu et al., 2007).

In recent years, the research focus has shifted from the importance of password policy provision to finding the best password policy to implement in PCSs to ensure that usable and strong passwords are created. Researchers have paid close attention to answering this question because the current guidelines are based on theoretical estimates (Burr, Dodson, & Polk, 2004) and not empirical data. Kelley et al. (2012) and Komanduri et al. (2011) found that usable and strong passwords are created by enforcing a password policy requiring a minimum of 16 characters without any further restrictions, such as including different character classes. However, Shay et al. (2014) argue that having only a length requirement makes users create passwords that are very easy to guess (e.g. *passwordpassword*), which negatively affects the security of the system. In their study, they examined various password policies that varied in the required length and character classes. Their results show that a password policy that enforces at least 12 characters including at least three different character classes has both usability and security benefits.

2.4.2.2 Password Creation Suggestions

Password creation suggestions are another supporting feature often provided by PCSs to help people create stronger passwords. However, some PCSs do not offer such suggestions because it is believed that they pose a risk to the security of systems. In contrast, Adams and Sasse (1999) assert that when users are not guided on password content, they come up with their own ideas and rules that lead to insecure passwords. Therefore, they recommend offering suggestions for password creation, but they do not specify what a good suggestion might be.

Two approaches are generally investigated in the research literature regarding how to advise users to create memorable and secure passwords. One is the mnemonic *phrase-based passwords* approach, and the other is the use of *chunking*. The majority of studies have focused on the former. A few studies have also been conducted on other approaches, but none have been implemented or recommended in practice.

The mnemonic *phrase-based passwords* approach was first proposed by Barton and Barton (1984), who advised users to choose the first letter of words in a phrase they found memorable. For example, the password could be ‘MbNi18yo’ for the phrase ‘My brother Nizar is 18 years old’. In this case, the password has a meaning to the user without revealing any explicit information about the password. Yan et al. (2004) conducted the first experiment on password advice with the aim of determining a helpful way to choose secure passwords. They studied the effect of giving different types of password creation advice on password creation and recall, and found that phrase-based passwords were as memorable as user-chosen passwords and as secure as randomly chosen passwords. However, evidence from Kuo, Romanosky, and Cranor (2006) suggests that mnemonic phrase-based passwords may not be as secure as previously assumed, since users tend to base their passwords on common phrases that are easily found on the Internet even though there are many different ways to generate such a password (Vu et al., 2007).

The second approach is based on *chunking* theory (Cowan, 2001; G. A. Miller, 1956). Carstens, Malone, and McCauley-Bell (2006) used chunking to help users create

memorable passwords. An example of a two-chunk password is ‘Rs#08-2193’, where the first chunk ‘Rs’ stands for Ryan Smith, and the second chunk ‘#08-2193’ stands for August 21, 1993. In their study, the authors investigated the effect of different formations of chunking passwords on password creation and recall. Their results showed that a four-chunk password was more memorable and longer than user-chosen passwords. However, the chunk-based passwords could be insecure since the participants were told what to use for their chunks.

2.4.2.3 Password Strength Indicators

The password strength indicator is the third possible supporting feature in PCSs. It is used to estimate the password strength to encourage users create stronger passwords, and typically comprises a two-dimensional visual representation of password strength (also known as password meter). Some current PCSs have implemented this type of strength indicator, such as Gmail, Twitter, and eBay. Recently, Kafas, Aljaffan, and Li (2013) proposed a new scheme for a password strength indicator, called the *visual password checker*. The key feature of the visual password checker is that it gives users feedback on the multiple threats to which their chosen password might be vulnerable.

In the literature, the majority of studies have focused on the algorithms used in password strength indicators (e.g. Castelluccia, Dürmuth, & Perito, 2012) instead of on the success of these indicators in creating strong passwords. However, a few recent studies have paid attention to the design aspect of these indicators and their effects on password creation and recall, such as the studies by Ur et al. (2012), Egelman et al. (2013), Vance et al. (2013), Khern-am-nuai et al. (2017) and Furnell and Esmael (2017).

Ur et al. (2012) analysed the effect of different variations of password indicators by evaluating 14 designs. The results showed that using any password strength indicator resulted in passwords that were more difficult to guess than those obtained without a password strength indicator. Furthermore, the authors found that password strength indicators led to longer passwords on average. Regarding memorability, there was no significant difference between different indicators. However, users were annoyed by

stringent indicators, which may have caused the users to ignore them. The authors also identified important features to consider when designing password strength indicators: a stringent scoring system and a visual component were the most important features for a good design, whereas colour, segmentation, and size were the least important. Furnell and Esmael (2017) further support the findings of Ur et al. (2012) that the mere presence of a password strength indicator increases the strength of password.

Egelman et al. (2013) found that password strength indicators only influenced password strength when the password was associated with a high-risk account. Thus, the password creation behaviour is dependent on the context in which the password is to be used. On the other hand, Vance et al. (2013) and Khern-am-nuai et al. (2017) found that the effectiveness of the password strength indicator depends on the information presented by the indicators. Severity, threats' susceptibility or warning messages are examples of the information that should be combined with the password strength indicators.

2.4.3 Designing Password Creation Systems

Although PCSs and their supporting features have existed for long time, little research has evaluated existing PCSs as whole interactive systems in their own right and the usability of the support they provide to users (e.g. Conlan & Tarasewich, 2006; Furnell, 2007; 2011). However, such an evaluation would help in understanding how these strategies could better influence the password creation and recall process. Evidence from Petrie and Power (2012) showed that users encounter usability problems when creating passwords. When users struggle to understand how a PCS works, this absorbs their cognitive effort, which could be used to create a strong and yet memorable password. A recent study has shown that cognitive effort is necessary for creating good passwords (Groß et al., 2016).

A preliminary report by Conlan and Tarasewich (2006) analysed a number of PCSs in terms of usability principles. The authors outlined a study to evaluate four different PCSs they developed, but did not report any results. In another study, Furnell (2007) analysed the password practices of 10 popular websites. He examined the password

policy and creation suggestion statements, among other features (e.g. password recovery), by creating user accounts and passwords. The results were disappointing. The practices of the websites assessed varied significantly. In addition, the statements provided were not helpful and were often ambiguous. Furnell (2011) repeated the study four years later and found similar results: there was no improvement in the design and implementation of the PCSs on these popular websites. In his second study, he included further features to examine, including password strength indicators. The findings of this recent assessment revealed that 90% of the websites provided a statement of password policy; of those websites, all required a minimum password length but only 22.2% covered a character classes requirement. On the other hand, 50.0% of the sample provided password creation suggestions, and 60.0% used a password strength indicator in their PCSs.

2.5 User Behaviours Related to Password Creation and Management

The tendency to forget passwords and the need to maintain multiple passwords has become a concern in the field of usable security. Users cannot cope with the burdens placed on them by the many systems that they need to use, all of which seem to require a password.

2.5.1 Problems with Passwords

2.5.1.1 Number of Multiple Passwords

From a security point of view, users should increase the number of passwords they have as the number of accounts they hold increases. However, evidence from Gaw and Felten (2006) indicates that this is not the case. Their results show that users accumulate more accounts over the years but do not have significantly more unique passwords. Using a client's component on users' machines, a large-scale study of password habits conducted with half a million users over a three-month period showed that each person had on average 25 accounts (Florencio & Herley, 2007). That data was collected in 2006, so the figure is undoubtedly considerably higher now. Yet the

number of different passwords that people use does not appear to be appropriate for this number of accounts (Florencio & Herley, 2007).

The present author compiled results from seven studies on password behaviour (Brown et al., 2004; Dhamija & Perrig, 2000; Florencio & Herley, 2007; Grawemeyer & Johnson, 2011; Gredler, 2012; Ponemon Institute, 2006; Technologies, 2003) conducted between 2000 and 2012. Figure 2.6 illustrates the range of the number of passwords (minimum to maximum) per person reported in these studies.

These studies found that the average number of passwords is approximately five, and the figure does not seem to be increasing, while the Web and the number of password-protected systems have grown since 2000. Indeed, the average number of passwords that people use has remained at the level predicted by Adams and Sasse (in 1999). This means that people are undoubtedly reusing passwords and creating security risks, as evidenced by a number of studies (Brown et al., 2004; Florencio & Herley, 2007; Gaw & Felten, 2006; Shay et al., 2010).

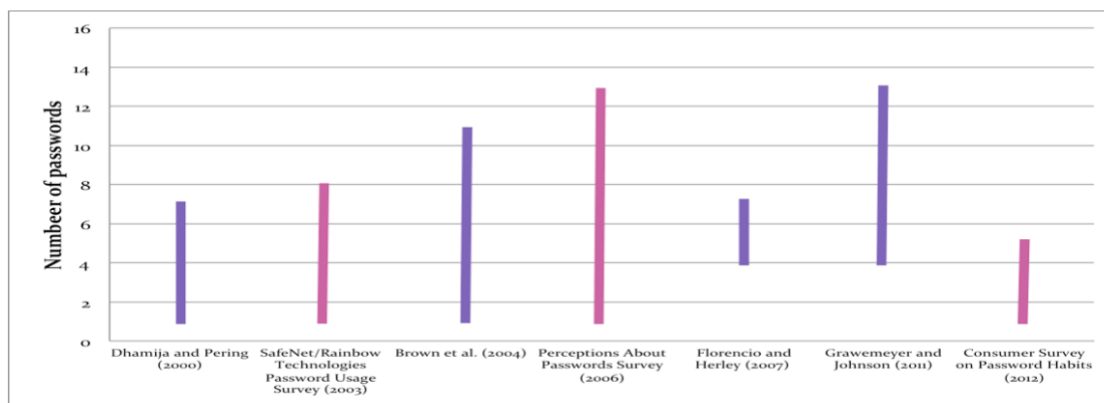


Figure 2.6 Minimum to maximum number of passwords held, across seven surveys (see text) conducted between 2000 – 2012

2.5.1.2 Forgetting Passwords

Another problem users face with the multiple passwords they have is forgetting passwords. A number of researchers have reported that users often forget their passwords due to human memory limitations, which consequently compromise security. A number of studies have demonstrated the existence of this problem (e.g.

Brown et al., 2004; Shay et al., 2010). For instance, Brown et al. (2004) surveyed 218 students, and 31.1% forgot their passwords at least once. Most of the passwords were for email, a university computer, or a phone card. However, a paper-based survey conducted by Shay et al. (2010) showed that out of 470 participants, only 19% of participants forgot their new password which they created as part of the study. Of this 19%, 60% remembered their password later, 21% retrieved it from a written note, and 11% contacted the help desk.

In addition, Shay et al. (2010) wondered whether there was a significant difference between the behaviours of individuals with different characteristics. Hence, they divided the participants according to their gender, age, occupational role (e.g. student, faculty, other university staff), and IT experience. The findings showed that role and gender could affect the forgetting problem. For example, faculty and staff were three times more likely to forget their passwords than students. Furthermore, women were twice as likely as men to forget their passwords. In contrast, forgetting was not related to IT experience or age. In a similar vein, Grawemeyer and Johnson (2011) conducted a diary study with 22 participants who had different occupational roles (e.g. administrative staff, researchers, lecturers, systems engineers, PhD and MSc students). They found that 52% of 48 unsuccessful password entries reported by the participants were related to memory. The reported memory failures included password misremembering, password interference, and forgetting the password altogether.

2.5.2 Coping Strategies

Coping strategies is a term used to refer to the strategies that users develop and use to cope with password problems. In 1999, Adams and Sasse were the first to document these coping strategies, and they are still in use since passwords continue to burden users' memories (Inglesant & Sasse, 2010).

In the literature, several methodologies have been used to investigate users' behaviours with PCs, including diary, interview, and questionnaire studies. In general, these studies have identified two common strategies: writing passwords down and reusing passwords. While other coping strategies are used, little evidence is

available about them. These include users being unlikely to change a password once it is set, the display of passwords in obvious locations, and sharing passwords with a third party.

2.5.2.1 Writing Down Passwords

The coping strategy of writing a password down can occur as a response to the demand of learning a new password, due to the requirement for a complex password or a policy of frequent password changes. Adams and Sasse (1999) first mentioned this behaviour in their study. The results showed that half of their 139 questionnaire respondents admitted to writing down their passwords. One of their questionnaire respondents stated, '*... because I was forced into changing it every month I had to write it down*'. Brown et al. (2004) and Dhamija and Perrig (2000) reported similar results to those of Adams and Sasse's study.

However, other studies have not reported users writing passwords down frequently, including those of Grawemeyer and Johnson (2011) and Shay et al. (2010). The reason for this may be linked with either the emergence of other coping behaviours, such as reusing the same passwords, or the absence of strict password policies, as indicated by Inglesant and Sasse (2010). Inglesant and Sasse (2010) investigated the password behaviours of participants in two organisations. They found that 9 out of 15 participants in organisation A wrote their passwords down, whereas none of the 17 participants in organisation B admitted to this strategy; the latter organisation's password policy allowed users to choose more memorable passwords.

2.5.2.2 Reusing Passwords

Another strategy practised by users to cope with the need to remember a large number of passwords for multiple accounts is the use of the same or a variation of existing passwords. Das et al. (2014) observed that 43.0% of users reuse the same password across multiple sites. Evidence from both controlled laboratory and field studies confirms the common practice of reusing passwords as a coping strategy.

In a laboratory study, Gaw and Felten (2006) asked participants to identify which websites they used, login to these websites, and write down their passwords. After finishing all logins, participants were asked to self-report the following: the number of passwords they used in the experiment, the number of unique passwords, the number of similar passwords, the number of password repetitions, and the number of passwords with related meanings. The results showed that participants reused on average three or fewer unique passwords in all their different passwords.

In another study, Brown et al. (2004) found that users had 8.18 passwords on average, whereas the average number of unique passwords was 4.45. Therefore, the average number of uses per password was 1.84. Out of 218 participants, 93.9% admitted to duplicating their passwords several times: once (37.4%), two times (29.7%), three times (20.9%), and four or five times (5.9%). However, users also sometimes varied one password for different accounts. In Shay et al.'s (2010) study, 80.0% of participants (out of 470) admitted to reusing their passwords, but 66.6% of these participants modified one password for different accounts.

In a field study, Florencio and Herley (2007) installed a client's component on users' machines to collect data about password habits. They found on average that each user had 25 accounts, but the average number of unique passwords was only seven. While the most likely reason for the widespread practice of reusing passwords is the burden of memory, the sensitivity of the account could also be a factor. For instance, Grawemeyer and Johnson (2011) found a significant positive relationship between the uniqueness of the password and the sensitivity of the account. In their study, the sensitivity of the account was based on users' self-assessment. The results also showed that out of 175 password-based accounts, 69 passwords were unique, 86 were completely reused in other accounts, and 20 were partially reused. This means that as the sensitivity of the account increases, the use of unique passwords increases and vice versa.

2.6 Password Security

Nowadays, there is a significant demand for high levels of security in authentication systems such as those described above. Any proposed new scheme should be as secure as possible, for example by having a large password space⁴ and considerable resistance to security attacks. It is therefore crucial to have a good understanding of the password strength and possible security threats of attempted authentication system break-ins.

2.6.1 Password Strength

Two methods can be used to measure password strength: password entropy and password guessability. The password entropy measure considers the password length and character composition (e.g. uppercase letters, lowercase letters, digits, and symbols) used in passwords. The password entropy provides a theoretical number representing how unpredictable a password is. The National Institute of Standards and Technology's published guidelines use different variations of entropies to measure password strength (Burr et al., 2004). On the other hand, the password guessability measure indicates the number of guesses required by password-cracking algorithms to guess the given password (Kelley et al., 2012).

2.6.2 Security Threats

There are two categories of security threat to any authentication system: malicious and non-malicious. The risks of these threats could be higher if the authentication method uses single-sign-on systems, which allows users to authenticate themselves through their social profiles (e.g. using their Facebook account). Using such a system could

⁴ The password space is the set of all passwords that it is possible to create for a given password policy. The larger the set of passwords, the greater the password space generated, which consequently increases the security by reducing the predictability of the passwords.

cause several issues for the users as well such as locking them out and compromising their accounts (Garfinkel & Lipford, 2014).

A malicious security threat is any electronic action taken by illegitimate users to exploit a system or user's digital identity with the intention of accessing the genuine user's data without permission (C. P. Pfleeger & Pfleeger, 2006). There are two classifications of malicious security threat: guessing attacks and capture attacks (Biddle et al., 2012).

Guessing attacks involve gaining access within a possible number of guesses by searching the password space exhaustively or predicting the likely password. These two methods are called exhaustive searching and dictionary attacks, respectively. In exhaustive searching, also referred to as brute-force searching, the attacker keeps searching the entire password space for possible valid passwords. Therefore, the possibility of resistance to this type of attack is more likely when the authentication system has a large password space. On the other hand, dictionary attacks involve searching a set of possible passwords looking for a match. The greatest defence against this type of attack is to impose policy restrictions when users choose their passwords, as there is a tendency to choose a weak password.

In contrast, capture attacks involve obtaining the password credentials, or part of them, by tricking the user or observing his or her login process. Phishing, shoulder-surfing, and social engineering are common malicious attacks against authentication systems. First, in phishing attacks, attackers attempt to trick the user with a fraudulent website to capture the user's password credentials. In social engineering, on the other hand, the attackers attempt to trick users by asking them to describe their password credentials verbally or in text. Finally, shoulder-surfing refers to an attacker looking over users' shoulders while they enter their password. Most researchers agree that shoulder-surfing is a more serious attack than the other attacks as it happens without users' awareness (Wiedenbeck, Waters, Sobrado, & Birget, 2006). However, Maguire and Renaud (2012) argued that shoulder-surfing is not an issue to worry about as it depends on the context of authentication.

Conversely, a non-malicious security threat exposes someone's digital identity to trustworthy persons such as family or friends without any expectation of harm. According to Adams and Sasse (1999), people tend to share passwords either because they have many passwords or because their passwords are too complicated. From the users' point of view, sharing passwords with trustworthy people can help them remember their passwords, but this practice is considered to be highly insecure. Nevertheless, there are some situations in which people have no other option but to share their passwords. For example, some disabled people rely on their relatives or carers to perform an authentication process for them.

2.7 Conclusions

In conclusion, users have difficulty choosing secure passwords and might not know how to do so. They choose passwords that are easy for them to remember and not at all difficult for password crackers to attack. Previous studies have focused on the strength and memorability of chosen passwords instead of looking at how supporting features are integrated into the user interface of PCSs and their effect on password choice. The usability of PCSs as whole interactive systems in their own right and the usability of their supporting features is one of the areas that has not been well studied. Therefore, it is important to further investigate the area of PCSs and improve the design and usability of these systems, ensure they effectively support users when they create passwords.

*Phase 1: Understanding the Current Practices of
PCSs*

Chapter 3

An Analysis of 30 Current PCSs – *Study 1*

3.1 Introduction

Password creation systems (PCSs) are small interactive web-based systems that are incorporated into most password-protected websites. Most users interact with PCSs when they sign up for a website during the registration phase or when they reset their password; this study focuses on the former case. Despite the fact that these systems have existed for some time, PCSs usually still have two main problems in their design: the lack of guidance to users during the password creation process, and the lack of consistency between different systems, which, it can be assumed, may negatively affect users' behaviour when creating passwords.

Furnell (2007) examined 10 popular websites in terms of their passwords practices. He looked at three key aspects of the PCSs: restrictions and policies to avoid weak passwords, suggestions for creating good passwords, and the implementation of the password reset feature. The results were disappointing: the websites showed substantial variability in their practices, and the advice provided was not helpful and often ambiguous. Four years later, he repeated the same study and found similar results (Furnell, 2011). There was no improvement in the design and implementation of the PCSs on these popular websites.

Building on Furnell's work, this study aims to provide a better understanding of the password creation process by analysing the current practices of 30 PCSs, with a focus on these systems' components and structure. Understanding the current practices of PCSs is important to the improvement of these systems and consequently the quality of passwords created with them. To this end, two research questions are formulated:

RQ1. What do current PCSs typically consist of?

RQ2. How are current PCSs structured?

3.2 Method

Thirty PCSs were selected for the analysis. The first 24 were selected from the top global 100 entries on Alexa⁵ on 26 May 2014 of the websites on which they appeared. Additional criteria for inclusion were that (a) the PCS should be in English; (b) the website should have a dedicated PCS (i.e. not use Google or Facebook for log-in); and (c) the PCS should not generate passwords automatically for users. A further six PCSs were then included by the author for having interesting features that were not found in the top visited websites. For example, the six PCSs provided different design features for the password strength indicator, such as colour-coding scheme. Table 3.1 lists all PCS websites, along with their domains and their Alexa rating, where appropriate.

Each PCS was thoroughly examined for features and structure by the author and her supervisor. Many different passwords were tried on each PCS to elicit a wide range of behaviour. Notes were made regarding when and how the password policy of the PCS was presented to the user, what suggestions for good passwords or tips on the password creation process were provided, and how a password strength indicator (if provided) was deployed.

Table 3.1 PCSs analysed for this study

Website	Domain	Alexa Rating
Adcash	Online advertising	56
Adobe	Computer software company	69
Amazon	Online retailer	9
Apple	Online retailer	33
BBC	Online newspaper	62

⁵ Alexa: <http://www.alexa.com/topsites>

ClearBooks	Online accounting	N/A
CNN	Online newspaper	63
DailyMail	Online newspaper	90
Dropbox	File hosting	82
eBay	Online auction site	28
Facebook	Social networking	2
Go Daddy	Web hosting service	67
Google	Search engine	1
Gravatar	Globally recognized avatars	N/A
HCII	Management system for international conference	N/A
IEEE	Digital library	N/A
Imgur	Online image hosting	44
LinkedIn	Business-oriented social networking services	10
MSN	Web portal	35
Netflix	Internet streaming media	78
PayPal	Payment and money transfer service	34
Pinterest	Social networking	23
Springer	Digital library	N/A
Stackoverflow	Question and answer service for programmers	50
Tumblr	Micro-blogging platform and social networking	40
Twitter	Social networking and micro-blogging	7
University of York (UoY)	Staff training website for the University of York	N/A
Wikipedia	Online free encyclopaedia	6
WordPress	Blog web hosting	25
Yahoo	Search engine	4

3.3 Results

This section examines the components and structures of the 30 PCSs.

3.3.1 Components of the 30 PCSs

The 30 PCSs included three key features to help users choose passwords (see Figure 3.1). A statement of the password policy (see Figure 3.1a), suggestions for creating good passwords (Figure 3.1b), and a password strength indicator - also known as a password strength meter (Figure 3.1c). They sometimes also provide feedback to users

about weak passwords or violations of the password policy in their proposed passwords. The following presents each one of the three key features.

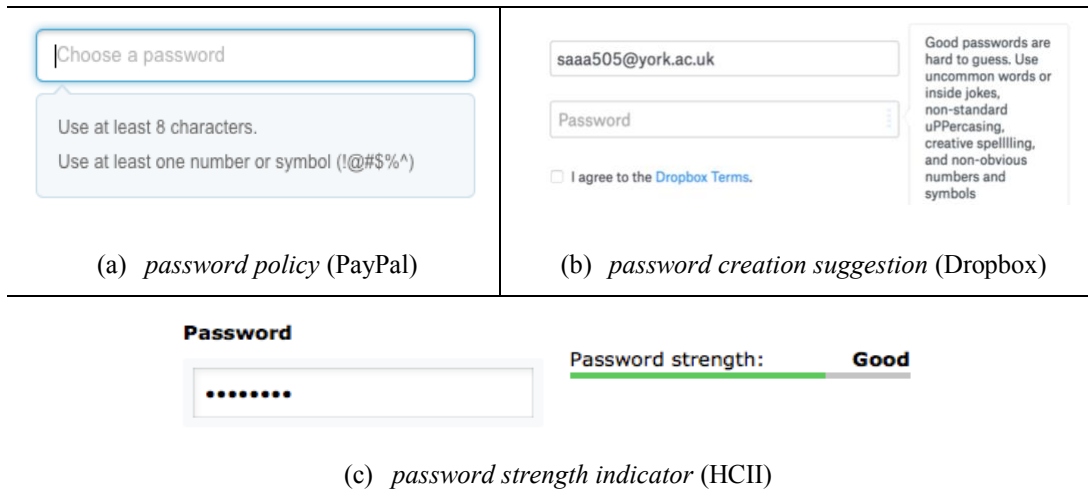


Figure 3.1 Examples of the three key supporting features in PCSs

3.3.1.1 Password Policies

Password policies are sets of rules that determine the accepted content of passwords for a given system. They typically include password length and types of characters that must (not) be included. Out of the 30 PCSs, 27 (90.00%) provided password policies at some point during the password creation process. An example of such a policy is, ‘*Password must have at least 6 characters and contain at least two of the following: uppercase letters, lowercase letters, numbers and symbols*’ (MSN). A closer analysis of the password policies identified three key attributes of interest: *language*, *content*, and *presentation*. The *language* attribute refers to the way the password policies are stated in the PCS. The second attribute, *content*, refers to the requirements for valid passwords. Thirdly, the *presentation* attribute refers to the way in which password policies are presented to users during the password creation process.

Table 3.2 shows the frequency of the occurrence of the types of language used in the password policy statements with the PCSs. Two forms of wording were identified in regard to the *language* attribute: password-oriented and action-oriented. In password-oriented language, the statement is related to the password, whereas in action-oriented

language, the statement is related to action that users should (not) take in creating their passwords. Almost 60% of the PCSs (59.26%) stated the policy using the password-oriented style.

Table 3.2 Frequency of the two language attributes used in the password policy ($N = 27$)

Attribute	Frequency (%)	Example	PCSs which provide this type of policy
Password-oriented	16 (59.26)	'Password must be at least 6 characters', WordPress	ClearBooks, DailyMail, IEEE, GoDaddy, Twitter, WordPress, Springer, LinkedIn, Apple, Tumblr, Imgur, Stackoverflow, Adcash, Facebook, BBC, UoY, Adobe
Action-oriented	11 (40.74)	'Use 6 or more characters', BBC	Amazon, eBay, Google, Yahoo, PayPal, Pinterest, BBC, CNN, Netflix, MSN, HCII

Table 3.3 shows the frequency of the content characteristics used in the password policy statements with the PCSs where they were encountered. In terms of the *content* attribute of the password policy statements, information was given about length (i.e. the minimum and/or maximum number of characters required in the password), the different character classes required (i.e. digits, case-sensitive letters, and symbols), and any exclusions of the use of special characters.

The 27 PCSs which offered password policies provided information to users about password length during the creation process; however, their practices were varied. For example, the majority of PCSs (74.07%) specified only the minimum number of characters, whereas almost a third of PCSs (29.63%) presented both a minimum and maximum number. Only one PCS stated a maximum length, but only when the candidate password exceeded this length.

Regarding the different character classes, 40.74% of PCSs presented a requirement of digits, 37.04% case-sensitive letters, or 22.22% symbols. However, some provided examples of these character classes within the password policy statements, while others did not. Finally, very few PCSs (7.40%) indicated exclusions of special characters in their policy statements.

Table 3.3 Frequency of content attributes used in the password policy statements ($N = 27$)

Attribute	Frequency (%)	Example	PCSs which provide this type of policy	
Length	Minimum	20 (74.07)	'Minimum 8 characters', ClearBooks Google, Facebook, MSN, BBC, HCII, Apple, eBay, Adobe, Amazon, GoDaddy, Imgur, Pinterest, Twitter, WordPress, PayPal, LinkedIn, Adcash, Tumblr, Stackoverflow, ClearBooks,	
	Maximum	1 (3.70)	'Your password can't be longer than 16 characters', MSN MSN	
	Both	8 (29.63)	'Your password must be between 4 and 10 characters', UoY Yahoo, BBC, CNN, Netflix, DailyMail, IEEE, Springer, UoY	
Digits	11 (40.74)	'Password must have at least one number', Apple Apple	eBay, MSN, Yahoo, IEEE, Apple, Springer, GoDaddy, PayPal, UoY, Stackoverflow, HCII	
Letters	With examples	3 (11.11)	'The password must contain at least one digit and one character (a-z)', Springer Springer	eBay, Springer, UoY
	Without examples	7 (25.93)	'Please enter a case-sensitive password', Amazon Amazon	Amazon, MSN, GoDaddy, Yahoo, Apple, Stackoverflow, HCII
Symbols	With examples	3 (11.11)	'Use at least one number or symbol (!@#\$\$%^)', PayPal PayPal	PayPal, UoY
	Without examples	3 (11.11)	'Use a mix of at least 6 letters (A-Z, a-z), numbers or special characters', eBay eBay	eBay, Stackoverflow, MSN
Exclusions	2 (7.40)	'...no spaces', CNN CNN	CNN, GoDaddy	

Table 3.4 shows the frequency of the presentation attributes used in the password policy statements with the PCSs where they were encountered. Three ways of presentation were identified: free text, bullet point, or inside password entry field. The majority of the PCSs (70.37%) presented their policy statements as free text. A quarter of PCSs (25.93%) provided their statements in a bullet-point list, which was combined with an indicator to help users determine which criteria were met and which were not

(e.g. using check/cross marks). Finally, one PCS presented its policy statement inside the password entry input field.

Table 3.4 Frequency of presentation attributes used in the password policy statements ($N = 27$)

Attribute	Frequency (%)	Example	PCSs which provide this type of policy
Free text	19 (70.37)	<i>'6 characters or more!'</i> , Twitter	Amazon, eBay, Google, IEEE, CNN, BBC, Twitter, WordPress, Springer, DailyMail, Imgur, Netflix, Facebook, Pinterest, Tumblr, Adcash, Adobe, LinkedIn, UoY
Bullet point	7 (25.93)	<i>'Please use: ✓ 8 to 32 characters ✕ upper and lowercase letters....'</i> , Yahoo	Yahoo, Apple, Stackoverflow, PayPal, GoDaddy, MSN, HCII
Inside password entry input field	1 (3.70)	<i>'Minimum 8 characters'</i> , ClearBooks	ClearBooks

3.3.1.2 Password Creation Suggestions

Password creation suggestions advise users on the content and structure of good and secure passwords. They may also offer more general advice about password behaviour, such as avoiding using the same password for multiple accounts. Out of the 30 PCSs, only 11 (36.67%) provided such suggestions during the password creation process. An example of a password content suggestion is *'Don't be afraid to use symbols like !%^£ (along with the numbers and letters'* (Gravatar), whereas an example of a password structure suggestion is *'Use uncommon words or inside jokes, non-standard uPPercasing, creative spelling and non-obvious numbers and symbols'* (Dropbox). Similar to the password policies, the same three key attributes were identified: *language, content, and presentation*.

Table 3.5 shows the frequency of the two possibilities for the language attribute used in the password creation suggestions for the 11 PCSs which included suggestions. Regarding the language attribute, 90.90% of these PCSs offered suggestions in an action-oriented format.

Table 3.5 Frequency of language attributes used in the creation suggestion statements ($N = 11$)

Attribute	Frequency (%)	Example	PCSs which provide this type of suggestion
Password-oriented	1 (9.09)	'Good passwords are hard to guess... ', Dropbox	Dropbox
Action-oriented	10 (90.90)	'Avoid using passwords you use for other sites ', eBay	Google, Pinterest, eBay, Twitter, WordPress, BBC, DailyMail, IEEE, UoY, Gravatar

Table 3.6 Frequency of content attributes used in the creation suggestion statements ($N = 11$)

Attribute	Frequency (%)	Example	PCSs which provide this type of suggestion
General	3 (27.27)	'Password could be more secure ', Twitter	Twitter, Pinterest, IEEE
Specific			
Abstract	3 (27.27)	'Don't use a password from another site, or something too obvious like your pet's name ', Google	Google, eBay, Dropbox,
Concrete	5 (45.45)	'Your password should be hard for anyone to guess, so we'd suggest using a mix of capitals, lower case and numbers for the strongest security ', DailyMail	WordPress, BBC, DailyMail, UoY, Gravatar

Table 3.6 shows the frequency of the content characteristics used in the password creation suggestions with the PCSs where they were encountered. In terms of *content* attribute, PCSs included both general and specific suggestions. The general suggestions provide users with a very broad statement about making stronger passwords without giving details on how to do so, unlike the specific suggestions which provide a detail advice. Of the specific suggestions, the suggestions were either abstract (in the form of an advice statement) or concrete (in the form of a requirement statement). Almost half of PCSs (45.45%) provided a specific concrete suggestion.

Table 3.7 shows the frequency of the presentation attributes used in the password creation suggestions with the PCSs where they were encountered. It shows that the majority of the PCSs (72.73%) presented their suggestion statements as free text. Only 27.27% of PCSs used a bullet-point list.

Table 3.7 Frequency of presentation attributes used in the creation suggestion statements ($N = 11$)

Attribute	Frequency (%)	Example	PCSs which provide this type of suggestion
Free text	8 (72.73)	'Great passwords use upper and lower case characters, numbers, and symbols like !"£\$%&', WordPress	Google, Twitter, Pinterest, WordPress, Dropbox, DailyMail, Gravatar, IEEE
Bullet point	3 (27.27)	'*Do not write your password down * Do not choose common phrases such as family names *Do not share your password', UoY	eBay, BBC, UoY

3.3.1.3 Password Strength Indicators

Password strength indicators are used to provide immediate feedback to users on the weakness or strength of their proposed password. Out of the 30 PCSs, 9 (30.00%) provided strength indicators during the password creation process. There are a number of different possible designs of these strength indicators. Two key design attributes were identified from the analysis: *media* and *colour schemes*. The *media* attribute refers to the media interface used to present the strength indicators, while *colour scheme* refers to the colours used to indicate the strength of the candidate passwords.

Table 3.8 Frequency of media used in the password strength indicators ($N = 9$)



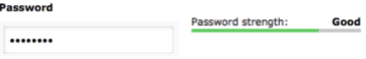
Attribute	Frequency (%)	Example	PCSs which provide this type of indicator
Graphical only	3 (33.33)	 , Twitter	IEEE, Twitter, Dropbox
Textual only	1 (11.11)	 , DailyMail	DailyMail
Graphical and textual	5 (55.56)	 , HCII	HCII, ClearBooks, eBay, Google, Apple

Table 3.8 shows the frequency of the media used in the password strength indicators for the 9 PCSs where they were used. 44.44% of PCSs used a single-medium indicator (graphic only or text only), whereas 55.56% PCSs used multimedia indicator (a combination of graphics and text). The graphical representations used were typically continuous progress bars or segmented bars, which present password strength in a

series of discrete steps. Regarding text, a set of words was used to indicate password strength, such as *weak*, *medium*, and *strong*.

PCSs also used colour to indicate the strength of the proposed passwords. Table 3.9 shows the frequency of the colour scheme attributes used in the password strength indicators with the PCSs where they were encountered. Both multi-colour and single-colour coding schemes were encountered in the strength indicators, with the majority of the PCSs (66.67%) using the latter. Some of the PCSs applied the “traffic light” metaphor. On the other hand, very few PCSs (33.33%) provided a single-colour indicator; these used either blue or green.

In general, the strength scale used varies greatly from one PCS to another, both in number of levels and the wording used to indicate those levels, ranging from three level (e.g. weak, medium, strong) to five (e.g. invalid, insecure, weak, OK, strong) levels.

Table 3.9 Frequency of colour scheme used in the password strength indicators ($N = 9$)

Attribute	Frequency (%)	Example	PCSs which provide this type of indicator
Multi-colour	6 (66.67)	‘ <i>Red, Amber, Green</i> ’, Apple/ DailyMail	Google, eBay, Apple, DailyMail, IEEE, ClearBooks
		‘ <i>Red, Amber, Green, Blue</i> ’, IEEE	
		‘ <i>Red, Amber, Yellow, Green</i> ’, eBay	
		‘ <i>Red, Amber, Blue, Green</i> ’, Google/ ClearBooks	
Single-colour	3 (33.33)	‘ <i>Green</i> ’, HCII & Twitter	HCII, Twitter, Dropbox
		‘ <i>Blue</i> ’, Dropbox	

3.3.2 Structure of the 30 PCSs

Although only three supporting features might be incorporated in PCSs, system providers have different approaches to implementing and using these features. These approaches differ in terms of *timing of presentation* and *provision of the supporting features*.

3.3.2.1 Timing of Presentation

The majority of PCSs presented their supporting features at different points during the password creation process. For example, some provided their password policy only when that policy was violated, while others informed users about their policy at the very beginning, even before a password could be entered. In contrast, some systems presented the policy dynamically while users entered their proposed password. Moreover, it is notable that some systems presented the features to users using all three of these options. Hence, it is clear that there was no standard point at which PCSs present supporting features to users. In this thesis, these different points of presentation are referred to as the timing of presentation. However, it is worth noting that they do not mean duration; they mean the sequence in which the supporting features are presented.

The analysis of the timing of presentation of the current PCSs led to the conceptualization of the password creation process as three-step model (see Figure 3.2):

- (1) *before-interaction* (Step 1) is the initial presentation before the users start to create a password, so when they open the page containing the password field. A password policy or suggestions for how to create good passwords may be presented in this step. Information presented at this step, the policy or good password suggestions may remain visible.
- (2) *during-interaction* (Step 2) is when the password entry occurs. In this step, information may be presented dynamically about the strength or appropriateness of the password as it is entered. Information presented at the first step, the policy or good password suggestions may remain visible or may be removed.
- (3) *after-interaction* (Step 3) is the step after the user has completed entering their password. At this step, error messages will appear if the password does not meet the policy and feedback may be given about the strength of the password.

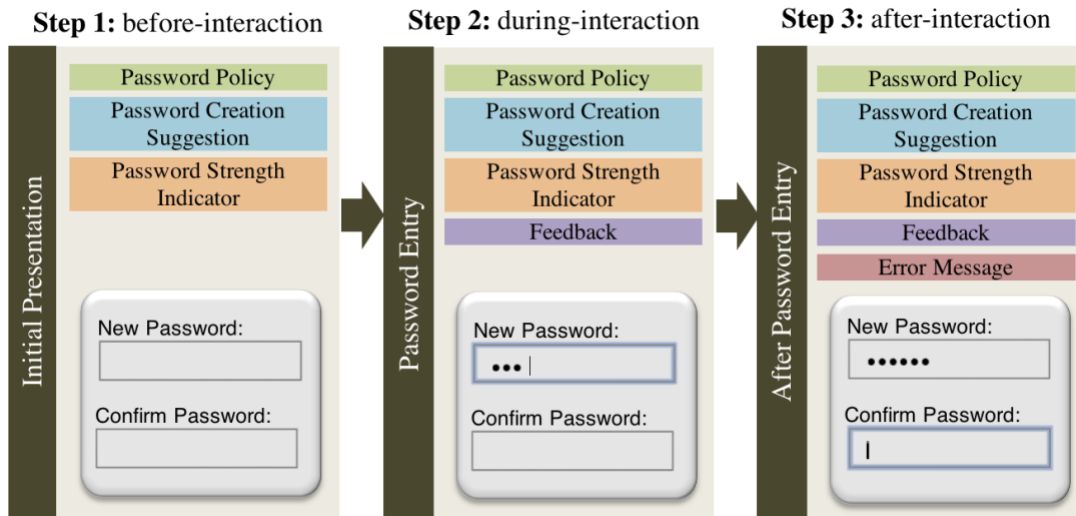


Figure 3.2 The three-step model of PCSs

Figure 3.3 illustrates the frequency of the three key features observed in the PCSs (policy, suggestions and strength indicators) in the three steps. Password policies, appeared with almost the same frequencies across the three steps. On the other hand, the highest occurrence of creation suggestions was in Step 1, before password entry. Finally, the majority of the strength indicators were provided dynamically while users entered a proposed password into the system (Step 2).

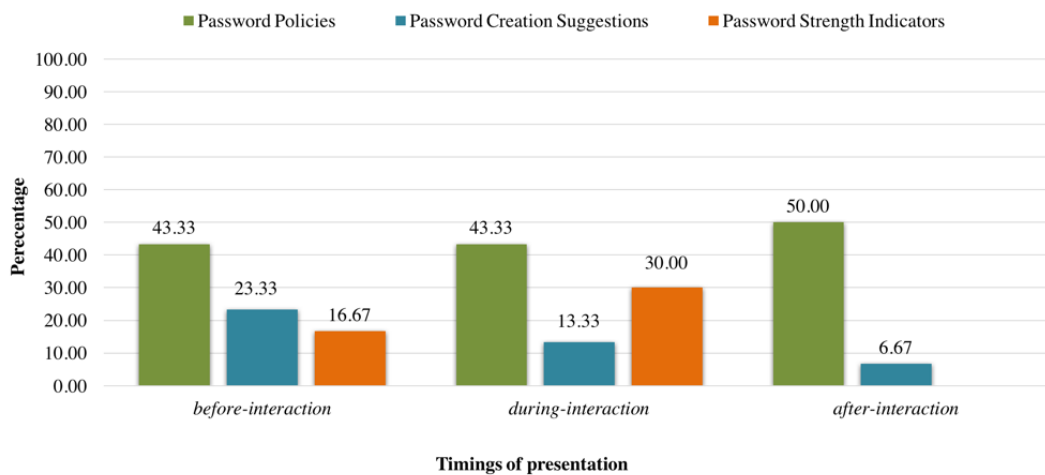


Figure 3.3 Temporal organisation of the three key features of PCSs (% total more than 100, as features can occur at more than one step in a PCS)

3.3.2.2 Provision of Supporting Features

Table 3.10 shows the frequency of the supporting features in terms of their provision and possible combination with the PCSs where they were encountered. The overall findings reveal that more than half of the analysed PCS provided one supporting feature (53.33%) during the password creation process, followed by two supporting features (26.67%), and all three of them (16.67%). Only one PCS out of the 30 did not provide any kind of help during the password creation process, this was Wikipedia.

As shown in Table 3.10, almost all PCSs that provided only one supporting feature (93.75%) opted to present their password policy instead of other features. Furthermore, presenting the password policy combined with either password creation suggestions or password strength indicators was more common practice than combining the latter two features. Although the provision of supporting features varied across the PCSs analysed, it is clear that providing the password policy during the password creation process is very common.

Table 3.10 Frequency of the provision of supporting feature ($N = 30$)

Provision of supporting features	Frequency (%)	PCSs which provide this type of supporting features
Never	1 (3.33)	Wikipedia
One supporting feature	16 (53.33)	
<i>policies</i>	15 (93.75)	Facebook, Yahoo, Amazon, LinkedIn, MSN, PayPal, Tumblr, Imgur, Stackoverflow, Adcash, CNN, Adobe, Netflix, Godaddy, Springer
<i>creation suggestions</i>	1 (6.25)	Gravatar
<i>strength indicators</i>	-	
Two supporting features	8 (26.67)	
<i>policies & creation suggestions</i>	4 (50.00)	Pinterest, WordPress, BBC, UoY
<i>policies & strength indicators</i>	3 (37.50)	Apple, HCII, ClearBooks
<i>creation suggestions & strength indicators</i>	1 (12.50)	Dropbox
Three supporting features	5 (16.67)	Google, Twitter, eBay, DailyMail, IEEE

3.4 Discussion

The present study aimed to develop an understanding of the password creation process. To address this aim, an analysis of 30 PCSs was conducted to assess their current practices in terms of components and structure. The first research question (RQ1) concerned the components of PCSs found in current practices, while the second (RQ2) related to the structure of these components in current PCSs. These research questions are answered in Section 3.4.1 and Section 3.4.2, respectively.

3.4.1 Components of Current PCSs

Three key features were used in current PCSs to help users choose passwords: a statement of password policy, suggestions for creating good passwords, and a password strength indicator. The results showed that the distribution of providing these supporting features varies greatly in current PCSs. Overall, the majority of the PCSs analysed (27/30, 90.00%) provided a password policy. But only a minority provided password creation suggestions (11/30, 36.67%) or password strength indicators (9/30, 30.00%). This means that more than half of the PCSs do not offer any suggestions about making strong passwords or provide a password strength indicator.

The occurrence of the supporting features in this study are compared to Furnell's (2011) findings. He assessed the enforcement of password restrictions (policy), provision of password guidance (suggestions), and use of password meter (strength indicators) on 10 popular websites. The 10 PCSs assessed in Furnell's study were included in the sample of this study. However, Furnell examined these features in both password creation (i.e. registration page) and password change (i.e. password reset page) stages. Therefore, the findings that related only to the password creation stage were compared since the current study focused only on the password creation stage. The findings of Furnell's assessment revealed that 90.00% (9/10) of the websites provided a statement of password policy. On the other hand, only 50.00% (5/10) of the sample provided a statement of creation suggestion, and 60.00% (6/10) used a password strength indicator in their PCSs.

Regarding the password policy, all of the PCSs in the current study and Furnell's study are matched in terms of the provision of the minimum length requirement. In both studies, the PCSs provided password policies with different character classes, yet with a different amount: whereas Furnell's study showed only two out of the nine PCSs having different character classes requirement, the current study showed an increase by 20% (12 out of the 27 policy statements). It is clear that progress is being made towards increasing the password space, making brute-force attacks less successful, and eventually creating stronger passwords.

Regarding the provision of creation suggestions, the frequency has decreased by 13%; in the current study (11/30, 36.67%) and Furnell's study (5/10, 50.00%). This finding may be explained by the fact that the current creation suggestions are often concerned about the passwords' constituent instead of advising users on how to create unpredictable/creative passwords.

For the provision of the password strength indicators, it was somewhat surprising to find a decrease in frequency of about 30%; in the current study (9/30, 30.00%) and Furnell's study (6/10, 60.00%). This could be due to the little consistency between the current deployment of commonly used indicators (Carnavalet & Mannan, 2015), which might lead to not implementing them in PCSs. For example, Furnell's study indicated that Yahoo provided a strength indicator during the password creation process, but this study did not find the same result (i.e. Yahoo has stopped providing a strength indicator).

In general, these findings are in line with Furnell's (2011) findings, which showed inconsistency and lack of guidance in providing the supporting features during the password creation process.

3.4.2 Structure of Current PCSs

With regard to structure, no prior research seems to have examined current PCSs in terms of their structure. This study proposed a three-step model of PCSs as interactive systems. The model conceptualizes the password creation process that users go through in terms of the steps and supporting features which may be available at each

step. Identifying these steps and supporting features of the PCSs may help to design and evaluate PCSs.

The findings showed that the provision of password policies did not differ by the timings of presentation; almost all statements were presented across the three timings of presentations. On the other hand, the provision of creation suggestions was very common before the password entry step. For the strength indicator, the majority of them were offered dynamically during the password entry step.

3.5 Conclusions

All in all, the findings of this study showed that the provision of supporting features and their organisation varies greatly. Most of the supporting features now available in PCSs may be implemented in an ad hoc manner, instead of considering what users want from these features, when the users want them to be implemented, and finally, how the users want them to be designed.

The current practices showed little consistency in designing and presenting these supporting features, and this may lead to user confusion and ambiguity as they cannot predict how a new PCS will work when they encounter it. Furthermore, the supported provided by existing systems does not seem adequate for users choosing a password. For example, some systems implement a password strength indicator as part of the PCS but do not offer any creation suggestions to increase password strength or explain why a chosen password is weak.

The question to address now is to what extent these current practices of PCSs could affect users in creating passwords. The current practices of PCSs may create usability issues for users and require too much cognitive effort, which can lead them to create weak passwords. To investigate the usability of PCSs further, it was decided to conduct usability evaluations. This is discussed in Studies 2 and 3 in Chapters 4 and 5, respectively.

Chapter 4

An Expert Evaluation of 12 Current PCSs – *Study 2*

4.1 Introduction

The analysis of PCSs presented in the previous chapter showed that there is little consistency in PCSs. This may create usability issues for users. To investigate the usability of PCSs further, it was decided to conduct usability evaluations on a number of current PCSs. In the first instance, an expert evaluation method was chosen as it provides a quick and inexpensive way of eliciting usability problems in interactive systems.

Very little research has been conducted on PCSs as whole interactive systems in and of themselves in terms of their usability: their effectiveness in supporting users in creating strong passwords, their efficiency in guiding users through the process quickly and without error without compromising security, and the users' satisfaction and experience. Only one paper could be found on this topic, a preliminary report by Conlan and Tarasewich (2006) who analysed a number of PCSs in terms of usability principles. The authors outlined a study to evaluate four different systems they developed themselves but did not report any results.

Numerous expert evaluation methods are used in human-computer interaction, including several varieties of cognitive walkthrough (Wharton, Bradford, Jeffries, & Franzke, 1992). However, the most commonly used method is heuristic evaluation (Nielsen, 1994). In heuristic evaluation, three to five usability experts individually work through an interactive system with a set of heuristics, identifying possible

usability problems and rating them for severity. The experts then come together and discuss all the problems, discarding any they cannot agree on as a usability problem and producing a final set of severity ratings.

Petrie and Buykx (2010) developed a variation of this method, collaborative heuristic evaluation (CHE), in which three to five usability experts work as a group throughout the evaluation instead of starting individually and then coming together for a discussion. The particular feature of CHE is that although any of the experts may propose a potential usability problem, and the group then discusses the precise nature of that potential problem, the experts then each rate the severity of the problem privately, and if they do not think it is in fact a problem, they simply rate it as a ‘non-problem’. At the end of the session, either the experts themselves or a facilitator create(s) a list of the agreed-upon problems, dropping any potential problems where too few experts have agreed it is actually a problem. The list also includes a measure of the severity of the problems perceived by the experts. In both classic heuristic evaluation and CHE, the experts use a set of heuristics of good usability principles to guide them through the evaluation.

In this study, a collaborative expert review was used. This is a general term referring to the inspection of a user interface by someone trained in usability without following a set of heuristics to test against the interface (Sauro, 2010). The same procedure of CHE was followed, but without the use of heuristics during the evaluation. Thus, experts were asked to draw on their general usability knowledge. Most of the well-known usability heuristics available were not suitable for evaluating PCSs: they were either too general or would overlook most of the details of interest in this research. However, one of the aims of this programme of research is to create a set of heuristics that could be used in future evaluations of PCSs. In addition, the experts were given the three-step model of PCSs (see Figure 3.2, Chapter 3) to guide them through each system. Therefore, the presents study aims to evaluate a number of PCSs using the expert method. The following research question were investigated:

RQ1. What usability problems does an expert evaluation identify in current PCSs?

4.2 Method

4.2.1 Design

Twelve PCSs were evaluated using a collaborative expert review method. Three to five experts worked together as a group evaluating three PCSs per session to identify usability problems and rate the severity of those problems privately on a four-point scale from ‘catastrophic’ to ‘cosmetic’. However, no set of heuristics was provided to the experts; instead, they were asked to draw on their general usability knowledge.

The experts were guided by the proposed three-step model of PCSs (see Section 3.3.2.1, Chapter 3). For each of the three steps in the model, the experts were asked in particular to identify any usability problems with each of three password creation support features: (1) password policies, (2) password creation suggestions, and (3) password strength indicators. In addition, they were asked to note any other usability problems that they thought were relevant to the password creation process.

The 12 PCSs were selected from the set analysed in Study 1 (see Section 3.2, Chapter 3). The 12 PCSs were chosen from the larger set based on the components they offered and the organisational structures of these components.

4.2.2 Experts

Seven usability experts participated in the study, all of whom worked or studied at the University of York. Three were women and four were men, and their ages ranged from 25 to 59 (mean 34.5 years). All had at least five years’ experience with usability evaluations, including expert evaluations. The experts were not compensated for their participation but were offered coffee and cookies during the evaluations.

Table 4.1 The 12 evaluated PCSs

PCS	Step	Supporting features		
		Password policies	Password creation suggestions	Password strength indicators
Amazon	<i>before-interaction</i>	✗	✗	✗
	<i>during-interaction</i>	✗	✗	✗
	<i>after-interaction</i>	✓	✗	✗
Apple	<i>before-interaction</i>	✗	✗	✗
	<i>during-interaction</i>	✓	✗	✓
	<i>after-interaction</i>	✓	✗	✗
DailyMail	<i>before-interaction</i>	✗	✗	✗
	<i>during-interaction</i>	✓	✓	✓
	<i>after-interaction</i>	✗	✗	✗
eBay	<i>before-interaction</i>	✓	✓	✓
	<i>during-interaction</i>	✓	✗	✓
	<i>after-interaction</i>	✗	✗	✗
Google	<i>before-interaction</i>	✗	✗	✗
	<i>during-interaction</i>	✓	✓	✓
	<i>after-interaction</i>	✗	✗	✗
MSN	<i>before-interaction</i>	✓	✗	✗
	<i>during-interaction</i>	✓	✗	✗
	<i>after-interaction</i>	✓	✗	✗
Netflix	<i>before-interaction</i>	✓	✗	✗
	<i>during-interaction</i>	✗	✗	✗
	<i>after-interaction</i>	✓	✗	✗
Stackoverflow	<i>before-interaction</i>	✗	✗	✗
	<i>during-interaction</i>	✓	✗	✗
	<i>after-interaction</i>	✗	✗	✗
Twitter	<i>before-interaction</i>	✓	✓	✗
	<i>during-interaction</i>	✓	✓	✓
	<i>after-interaction</i>	✗	✗	✗
Wikipedia	<i>before-interaction</i>	✗	✗	✗
	<i>during-interaction</i>	✗	✗	✗
	<i>after-interaction</i>	✗	✗	✗
WordPress	<i>before-interaction</i>	✗	✓	✗
	<i>during-interaction</i>	✓	✗	✗
	<i>after-interaction</i>	✓	✗	✗
Yahoo	<i>before-interaction</i>	✓	✗	✗
	<i>during-interaction</i>	✓	✗	✗
	<i>after-interaction</i>	✗	✗	✗
Overall	<i>before-interaction</i>	5	3	1
	<i>during-interaction</i>	9	3	5
	<i>after-interaction</i>	5	0	0

4.2.3 PCSs

The 12 PCSs used in the expert evaluation are listed in Table 4.1 along with their supporting features and their timing of presentation. In terms of supporting features, almost all 12 PCSs offered password policies; five provided password creation suggestions; and five presented password strength indicators. The combination of supporting features varied across the PCSs: four provided all three features and two provided two features. In addition, the chosen sample had a variability with respect to the timing of presenting these features. Regardless of the offered features, 17 instances were presented in the *during-interaction* step, followed by nine instances in the *before-interaction* and five instances in the *after-interaction* step.

For each PCS, a set of passwords was prepared for the experts, including valid and invalid passwords, and strong and weak passwords (see Appendix B, Section B.1 for the full set of provided passwords). The aim of providing a set of passwords was to show the strengths and weaknesses of the particular PCS; each set was based on the author's extensive exploration of the PCS. The experts were encouraged to try out the PCS with these passwords but were also free to try any other passwords to see their effects on the PCS.

4.2.4 Equipment and Materials

Two computers were used in the evaluations, and both were connected to projectors. One displayed the PCS to the experts, and the other the list of proposed usability problems.

Experts were given sheets with an outline of the PCS three-step model (see Appendix B, Section B.2), a summary of the severity rating schema, and sheets to record their private severity ratings. The experts were also given a short demographic questionnaire. This collected information about gender, age, qualifications, occupation, years of usability experience, and types of usability methods typically used, and opinions about the use of the three-step model during the evaluation sessions.

4.2.5 Pilot of the Study Procedure

A pilot study was conducted to evaluate the evaluation method by evaluating three PCSs (from the 12 PCSs) with three experts. The pilot took about approximately 1.5 hours. The experts found the proposed three-step model in the evaluation method easy to use. In addition, the overall procedure during the session was found to be easy. However, one of the experts noted that rating usability problems is difficult since the consequences of particular features of the PCS are not necessarily seen immediately and the real meaning of information is not known. In particular, how the PCS rates the strength of a password is not necessarily revealed. For example, some PCSs can accept a password even when it is rated as ‘weak’. On the other hand, other PCSs ask users to increase the strength of their password and try again if they enter a weak one. In the first case, a PCS that does not help users to create stronger passwords and would accept any passwords even weak ones is not considered to be a catastrophic problem, whereas it is in the second case. Therefore, in the main study, experts were given information about the enforcement policy for each PCS during the evaluation. The data from the pilot session was included in the data analysis.

4.2.6 Procedure

Four evaluation sessions were conducted with three to five experts participating in each, depending on their availability, different experts took part in different sessions. Each session lasted approximately two hours, with three PCSs being evaluated in each session. Each PCS took approximately 20 to 30 minutes to evaluate, allowing for short comfort breaks between PCS evaluations. Each session was led by a facilitator (the author or her supervisor).

At the beginning of the initial sessions, the facilitator introduced the aim of the study and briefed the experts on the procedure to be followed. One expert acted as ‘driver’ of the PCS, interacting with the system as requested by all remaining experts. The facilitator acted as scribe, recording the potential usability problems proposed by the experts, along with the step in the PCS model and the supporting feature that related to the usability problems. The PCS and the list of potential problems were displayed

on large screens via projectors so that the experts could view and discuss them easily. For each PCS, the experts completed only one task: creating a new password. However, they were asked to explore as many possibilities in this task as they wished to, trying out different passwords from the list provided and any others they wished, to see their effects in the PCS.

Any expert could propose a potential usability problem, and discussion was allowed about the precise nature of the problem. However, the experts were asked not to air their opinion publicly if they believed that it was not in fact a problem. When the description of the problem was agreed upon, each expert rated the severity of the problem privately using the four-point scale from the heuristic evaluation: 4 is a catastrophic problem, 3 is a major problem, 2 is a minor problem, and 1 is a cosmetic problem. If an expert did not think the potential problem was in fact a problem, he or she rated it as zero. Experts were also asked to privately note in which step in the model they thought the problem occurred and which supporting features were involved. This procedure was repeated until the experts felt there were no more problems to be identified in the PCS.

4.2.7 Data Analysis

A list was created of the usability problems identified by the experts during the evaluation sessions. In light of the three-step model, each usability problem was assigned to a supporting feature representing the content of the problem. The problems were then grouped in terms of these features. Thus, five main categories emerged: *password policies*, *password creation suggestions*, *password strength indicators*, *other feedback*, and *error messages*. A sixth category, '*other problems*', was added for usability problems that did not belong to any of the five main categories.

Subsequently, an open coding technique was used to sub-categorise the usability problems. The author and her supervisor repeated this process until appropriate category labels emerged. Then, the author invited a second coder to verify the coding categories to establish inter-code reliability. The second coder coded a random set of

13 problems. Cohen's Kappa (K) (J. Cohen, 1960) was used to measure the level of agreement between the coders. This was found to be excellent (K = 0.921).

4.3 Results

This section first examines the expert agreement on the existence of problem and their severity ratings. Then, it presents the number of usability problems in light of the three-step model. Finally, it demonstrates the categorisation of the usability problems.

4.3.1 Expert Agreement

A first issue to investigate was the level of agreement between experts on the existence problems and their severity ratings. As different experts took part in different sessions, pairwise comparisons were made between experts. Table 4.2 provides the overall statistics for all the experts. The number of potential problems rated by the different experts varied greatly, as different experts were available for different numbers of sessions and evaluated different number of PCSs. The number of times one expert proposed a usability problem and any of the other experts rated it as 'not a problem' was very low, occurring in less than 5% of cases for all evaluators, so this was not an issue.

Table 4.2 Overall statistics for each expert in the study

Expert	E1	E2	E3	E4	E5	E6	E7
Total no. of problems	100	84	71	118	60	45	12
Rated 'not a problem' (%)	4 (4%)	2 (2.38%)	1 (1.41%)	3 (2.54%)	2 (3.33%)	2 (4.44%)	0 (0%)
No. of PCSs evaluated	10	9	6	9	5	3	3
Mean rating of problems	2.30	1.80	1.97	1.97	2.24	2.09	1.75
Standard deviation	0.76	0.78	0.72	0.61	1.00	0.72	0.62

Table 4.3 summarises the pairwise comparisons between experts. There were 20 possible pairs, but only 15 pairs in actuality, as some experts did not participate in a session together. For 8 of the 15 pairs, there was a significant difference between the experts in their ratings, and for the remaining seven pairs there was no significant difference.

Table 4.3 Pairwise comparisons between experts on severity ratings of problems

Pair	No. of ratings in common	Mean rating (SD) first expert	Mean rating (SD) second expert	Z value	p value
E1 – E2	73	2.33 (0.77)	1.81 (0.79)	-5.48	.000
E1 – E3	66	2.30 (0.78)	1.98 (0.71)	-2.80	.005
E1 – E4	86	2.29 (0.77)	1.93 (0.61)	-3.57	.000
E1 – E5	30	2.27 (0.69)	2.03 (0.96)	-1.29	n.s.
E1 – E6	12	2.33 (0.78)	2.42 (0.79)	-0.38	n.s.
E1 – E7	8	2.63 (0.52)	1.75 (0.71)	-2.65	.008
E2 – E3	68	1.75 (0.78)	1.99 (0.72)	-2.16	.031
E2 – E4	69	1.72 (0.78)	1.88 (0.58)	-1.55	n.s.
E2 – E5	33	1.67 (0.78)	2.03 (0.95)	-2.03	.042
E2 – E7	11	2.27 (0.65)	1.73 (0.65)	-2.12	.034
E3 – E4	68	1.99 (0.72)	1.91 (0.57)	-0.86	n.s.
E3 – E5	33	2.27 (0.72)	2.03 (0.95)	-1.23	n.s.
E4 – E5	58	2.07 (0.56)	2.24 (1.00)	-1.15	n.s.
E4 – E6	42	2.12 (0.63)	2.07 (0.71)	-0.35	n.s.
E5 – E6	23	2.61 (0.99)	2.04 (0.71)	-2.50	.012

Note. Wilcoxon signed ranks test (Z) was used for the pairwise comparison.

Table 4.4 Cumulative percentage in levels of agreement between experts in severity ratings of problems

Pair	Exact agreement	Agreement + 1 difference	Agreement + 2 differences	Agreement + 3 differences
E1 – E2	42.5	97.3	100	100
E1 – E3	52.2	89.6	100	100
E1 – E4	47.7	88.4	100	100
E1 – E5	30.0	90.0	100	100
E1 – E6	41.7	100	100	100
E1 – E7	12.5	100	100	100
E2 – E3	48.5	89.7	100	100
E2 – E4	46.4	94.2	100	100
E2 – E5	42.4	84.8	100	100
E2 – E7	27.3	100	100	100
E3 – E4	58.8	92.6	100	100
E3 – E5	24.2	72.7	100	100
E4 – E5	24.1	87.9	98.2	100
E4 – E6	40.5	92.9	100	100
E5 – E6	21.7	86.9	100	100
Overall	37.4	91.1	99.9	100

However, the levels of differences between experts were not great. Table 4.4 shows the cumulative percentages in levels of agreements between pairs of experts. E1 and E2 gave exactly the same rating in 42.5% of instances, but were in agreement or differed by only one rating in 97.3%, and in agreement or differed by only one or two ratings in all instances. The overall mean level of agreement to within one rating was 91.1%, which seems highly acceptable given that individual experts will always have differences in their individual stringency in rating the severity of problems.

4.3.2 Usability Problems

In total, 131 usability problems were identified by at least one evaluator, with a mean of 10.92 per PCS (standard deviation = 6.32). Table 4.5 shows the number of usability problems per PCS as well as the mean severity rating of problems per PCS. The range of numbers of problems is large, from Apple with 23 problems to Amazon and Twitter with only four problems each. To some extent, it seems that the number of problems is dependent on the level of interaction of the PCS: the more features are offered at different timings of presentation, the more potential there is for problems.

Table 4.5 Total number of usability problems found by experts per PCS with severity ratings

PCS	Total no. of problems (%)	Mean severity rating (SD)
Amazon *	4 (3.05)	1.58 (0.10)
Apple	23 (17.56)	1.96 (0.59)
DailyMail	20 (15.27)	2.14 (0.49)
eBay	11 (8.40)	2.05 (0.38)
Google *	5 (3.82)	2.43 (0.37)
MSN	14 (10.69)	1.90 (0.40)
Netflix	7 (3.34)	1.54 (0.49)
Stackoverflow	17 (12.98)	2.20 (0.49)
Twitter *	4 (3.05)	2.25 (0.50)
Wikipedia	8 (6.11)	2.29 (0.65)
WordPress	8 (6.11)	1.89 (0.63)
Yahoo	10 (7.63)	2.19 (0.61)
Total	131	2.03 (0.46)

Note. * denotes the PCS that was evaluated in the pilot session.

Table 4.6 shows the number of usability problems found by experts at each step of the three-step model for each supporting feature. The results show that just of half of the usability problems (51.15%) occurred in the *during-interaction* step, where users start entering their proposed password. The other half of the usability problems were fairly equally split between the before and after steps (23.67% before and 25.19% after).

Table 4.6 Number of usability problems found by experts for each supporting feature and step

Supporting features	Steps (timing of presentation)		
	<i>before-interaction</i>	<i>during-interaction</i>	<i>after-interaction</i>
	Frequency (%)	Frequency (%)	Frequency (%)
<i>Password policies</i>	11 (35.48)	25 (37.31)	4 (12.12)
<i>Password creation suggestions</i>	12 (38.71)	7 (10.45)	3 (9.09)
<i>Password strength indicators</i>	0	13 (19.40)	1 (3.03)
<i>Other feedback</i>	0	11 (16.42)	8 (24.24)
<i>Error messages</i>	0	6 (8.96)	17 (51.52)
<i>Other problems</i>	8 (25.81)	5 (7.45)	0
Total	31 (23.67)	67 (51.15)	33 (25.19)

4.3.3 Categorisation of Usability Problems

Table 4.7 summarises the categorisation of the usability problems. Six main categories emerged: *password policies*, *password creation suggestions*, *password strength indicators*, *other feedback*, *error messages*, and *other problems*. In general, 38 subcategories were found, 10 of which occurred in the context of more than one main category. For example, the subcategory *information too detailed* occurred in both the *password policies* and *password creation suggestions* main categories.

As shown in Table 4.7, the *password creation suggestions* category was the most elaborated category with eight subcategories, whereas the *password strength indicators* category was the least elaborated one with six subcategories. The other four categories had seven subcategories each. Overall, nearly one-third of the usability problems (30.53%) were in the *password policies* category. In contrast, relatively few usability problems were found in the *password strength indicators* and *other problems* categories (10.69% and 9.92%, respectively). The remaining categories – *password*

creation suggestions, *other feedback*, and *error messages* – had relatively similar percentages of usability problems (16.79%, 14.50%, and 17.56%, respectively). In terms of the mean severity ratings of usability problems, the *other feedback* category had the highest mean severity rating, whereas the *other problems* category had the lowest mean severity rating.

Table 4.7 Number of usability problems identified by experts, with the number of PCSs in which they were encountered and mean severity ratings

Problem category	No. of problems (<i>N</i> =131)	No. of PCSs (<i>N</i> =12)	Mean severity (SD)	Example
<i>Password policies</i>	40 (30.53%)	11 (91.67%)	2.02 (0.49)	
1.1. Not provided or provided too late	8	7	2.14 (0.35)	No information provided about password policy.
1.2. Information not detailed enough	6	6	2.22 (0.53)	The non-alphanumeric characters listed – are these just examples or the only non-alphanumeric characters allowed?
1.3. Information too detailed	1	1	2.75	An overwhelming list of requirements.
1.4. Confusing/odd statement	23	9	1.87 (0.51)	3 bullet points for 4 requirements? Or do they mean numbers or symbols?
1.5. Inconsistency of terms	2	2	1.88 (0.18)	Symbols become special characters.
1.6. Querying the policy	1	1	1.25	Minimum length of characters is four, which is odd.
<i>Password creation suggestions</i>	22 (16.79%)	10 (83.33%)	2.11 (0.34)	
2.1. Not provided	8	7	2.06 (0.27)	No suggestions provided on how to create good passwords.
2.2. Information not detailed enough	4	3	1.81 (0.24)	No mention of using symbols, but creates strong password.
2.3. Information too detailed	1	1	2.00	Too much information on ‘great passwords’ page.
2.4. Unclear how to create a good password	2	2	2.25 (0.35)	Confusing that I have met all requirements but my password is only moderate [no more advice].
2.5. Unclear/confusing language	2	2	2.00 (0.47)	‘Strong security’ - will this mean anything?
2.6. Complex presentation	4	2	1.88 (0.37)	Suggestion boxes disappear when one might need them.
2.7. Querying the creation suggestions	1	1	2.75	PCS provides a password generator that creates passwords that are difficult to remember.
<i>Password strength indicators</i>	14 (10.69%)	7 (58.33%)	2.29 (0.29)	
3.1. Not provided	2	2	2.29 (0.06)	No indication of how strong a password actually is.
3.2. Poor colour contrast	3	2	2.06 (0.42)	Medium = orange, weak = red, very poor colour contrast.
3.3. Poor colour coding semantics	2	2	2.17 (0.24)	Red is not usually associated with something acceptable.
3.4. Poor/unclear/confusing presentation	2	2	2.50 (0.24)	What does the bar mean?
3.5. Timing issue	1	1	2.00	Message that password is too short/invalid occurs when cursor is still in password field – user may still enter more.
3.6. Querying the strength indicators	4	2	2.75 (0.32)	Why is it weak? I’ve met the requirements.

Problem category	No. of problems (N=131)	No. of PCSs (N=12)	Mean severity (SD)	Example
<i>Other feedback</i>	19 (14.50%)	10 (83.33%)	2.32 (0.37)	
4.1. Symbols not clear	4	4	1.87 (0.69)	What does the tick indicate? No explanation of it.
4.2. No feedback about valid/invalid password	5	5	2.67 (0.63)	Absence of confirmation that password meets policy – user has to infer this from lack of error messages.
4.3. No feedback about matching/non-matching password	3	3	2.53 (0.65)	No immediate feedback when password confirmation does not match.
4.4. Feedback not accurate	5	3	2.13 (0.46)	Caps lock message does not disappear when caps lock is turned off
4.5. Inconsistent presentation	1	1	2.00	Inconsistency in validity message: ‘invalid’ on the strength indicator but ‘✓’ when the password is valid.
4.6. Querying the security of the PCS	1	1	2.75	PCS accepts ‘Password’ as password and rejects ‘password’.
<i>Error messages</i>	23 (17.56%)	9 (75.00%)	2.13 (0.32)	
5.1. Information not specific enough	3	2	2.13 (0.12)	Violation messages do not clarify what to do for users.
5.2. Poor/unclear/confusing presentation	7	4	2.03 (0.52)	Error message is distant from the password – will a user associate them?
5.3. Inconsistent statement	4	3	1.99 (0.18)	Sometimes the message is about the structure or length versus the strength.
5.4. No visual distinction between error message and creation suggestions	2	2	2.67 (0.47)	The ‘password does not match’ message is too similar to suggestions – users may not realise it is a different message.
5.5. Timing issue	5	3	2.27 (0.72)	‘Password confirmation can’t be empty’ feedback comes as soon as password meets policy, but before the user has tried to confirm.
5.6. Unclear/confusing language	1	1	1.60	‘Your passwords don’t match’ suggests we have two different passwords, but we only one that we are verifying.
5.7. Querying the security of the PCS	1	1	2.25	Why is ‘mM123456’ rejected as a common password?
<i>Other problems</i>	13 (9.92%)	7 (58.34%)	1.34 (0.37)	
6.1. Unclear/confusing language	3	3	1.31 (0.34)	‘Choose a password’ is odd wording
6.2. Poor/unclear/confusing presentation	6	5	1.46 (0.51)	No explicit labels on the entry fields
6.3. Inconsistent presentation	1	1	2.00	Inconsistent presentation of the ‘this information is required’ message.
6.4. Poor colour contrast	1	1	1.25	Very low contrast on ‘bullet’.
6.5. Symbols not clear	1	1	1.00	Is this a bullet point or a radio button? What is it?
6.6. Functionality lacking	1	1	1.00	No confirm password field provided.

4.4 Discussion

The aim of this study was to conduct an expert evaluation of 12 current PCSs to investigate what usability problems they would identify and what severity ratings experts would give to those problems. Experts were asked to use the three-step model to guide them through the evaluation as it would help them focus more clearly and efficiently on the usability problems of the PCSs.

All the experts stated that the three-step model (also referred to as timing of presentation – see Section 3.3.2.1, Chapter 3) was easy to understand and helpful in the evaluation. For example, it helped them to structure their consideration of the context of possible usability problems: when the usability problem occurred and in which category of supporting features it belonged. The experts also appreciated that each PCS took only 20-30 minutes to evaluate but considerable numbers of usability problems were identified, which surprised them. The overall mean level of expert agreement to within one rating was 91.1%, which seems highly acceptable given that individual experts will always have differences in their stringency for rating the severity of problems.

The number of usability problems found by experts in the PCSs was surprisingly high, 131 in total, even though the PCSs are very small and seemingly simple interactive systems. Thus, developers of PCSs need to take more care regarding the usability of these systems and could benefit from support in development and usability evaluation of PCSs.

Using the three-step model, six main categories of usability problems emerged: the statement of *password policy*, the statement of *password creation suggestions*, *password strength indicators*, *feedback* to the user about the password creation process, *error messages* in case of password violations, and *other problems* for did not belong to any of the five main categories.

Interestingly, a set of patterns can be seen in the subcategories in each of the main categories: not providing enough information, providing too much information, and confusing statements and presentations. Furthermore, one subcategory of problems may not strictly be *usability* problems: the experts questioned why the PCS accepted an overly simple password or rejected a seemingly complex one. These issues were included in the analysis as they are further aspects of a PCS that may distract or confuse users, require cognitive effort, and thus lead to poor password creation.

Finally, the results showed that just half of the usability problems occurred during the second step of the model, during actual interaction with the system when users enter their password. This is not surprising for two reasons. Firstly, the interactive elements clearly provide the most possibilities for usability problems. Secondly, in this step, users must in effect split their attention between two tasks: creating a good password and understanding what is happening with the PCS. While providing dynamic feedback during the password entry step may seem like a good idea, this split attention issue needs to be considered.

4.5 Conclusions

All in all, the three-step model of PCSs and the evaluation of current PCSs highlighted the considerable number of usability problems with PCSs. The number of usability problems found by the experts was surprisingly high, 131 in total, even though PCSs are very small interactive systems. Therefore, it was crucial to investigate whether users would also encounter these problems using a user evaluation. This is discussed in Study 3 in Chapter 5.

Chapter 5

A User Evaluation of Six Current PCSs – *Study 3*

5.1 Introduction

It is important to complement expert evaluation with usability testing including potential or real users of systems (Shneiderman, Plaisant, Cohen, Jacobs, & Elmqvist, 2018) for several reasons. Firstly, expert evaluation relies on experts' knowledge and opinions, and different experts tend to find different sets of problems in an interactive systems or may have conflicting opinions (Hertzum & Jacobsen, 2001; Hertzum, Jacobsen, & Molich, 2002; Shneiderman et al., 2018). In addition, given the high number of usability problems found by experts in Study 2, it is crucial to investigate whether the problems found by the experts are also encountered by users, and to identify what types of problems users encounter in current PCSs. To address this aim, the following research questions are formulated for this study:

RQ1. What usability problems does a user evaluation reveal in current PCSs?

RQ2. Are there differences in the types and numbers of usability problems that expert and user evaluation methods find?

The first research question (RQ1) is addressed by evaluating six current PCSs with a widely used method for usability testing: a concurrent think-aloud protocol (Birns, Joffre, Leclerc, & Paulsen, 2002; Dumas & Redish, 1999; Van den Haak & de Jong, 2003). The second research question (RQ2) is addressed by comparing the usability problems found by the experts (Study 2, Chapter 4) and those identified by the users (Study 3, present chapter).

5.2 Method

5.2.1 Design

Six PCSs were evaluated using a concurrent think-aloud protocol. The think-aloud protocol involves participants speaking out their thoughts as they do specified tasks. One can conceptualise this protocol as participants giving researchers a running commentary on what they are doing and thinking. A within-participants design was used in this study. Each participant evaluated all six PCSs, undertaking one simple task per PCS: creating a new password. Following the think-aloud protocol, participants were asked to identify usability problems and rate their severity.

5.2.2 Participants

24 participants took part in the study. All of them worked or studied in the Department of Economic and Related Studies or the Department of Theatre, Film and Television at the University of York; 22 were university students and 2 were administrative staff. Of the 24 participants, 11 were women and 13 were men. The participants' ages ranged from 18 to 33 years, with a mean of 21.79 years (standard deviation = 3.96). The majority of participants (18, 75.00%) were native English speakers, and the remaining had spoken English for an average of 10.33 years (standard deviation = 4.08). When asked about the frequency of their daily use of the internet and any type of computer, they reported in the range of 'always' to 'often' on a five-point scale. No participant (except for one for one PCS) had created an account for any of the PCSs evaluated in the study in the last month. Participants were remunerated for their efforts with GBP 15 Amazon gift vouchers.

5.2.3 PCSs

Table 5.1 lists the six evaluated PCSs along with their supporting features and their timing of presentation. Six PCSs were selected from the set already evaluated by the experts (see Section 4.2.2, Chapter 4). The criteria for the chosen sample were: (1) the number of problems identified by experts, (2) the components that PCSs offered, and (3) the organisational structures of these components.

Table 5.1 The six evaluated PCSs

PCS	Step	Supporting features		
		Password policies	Password creation suggestions	Password strength indicators
Apple	<i>before-interaction</i>	✗	✗	✗
	<i>during-interaction</i>	✓	✗	✓
	<i>after-interaction</i>	✓	✗	✗
DailyMail	<i>before-interaction</i>	✗	✗	✗
	<i>during-interaction</i>	✓	✓	✓
	<i>after-interaction</i>	✗	✗	✗
Netflix	<i>before-interaction</i>	✓	✗	✗
	<i>during-interaction</i>	✗	✗	✗
	<i>after-interaction</i>	✓	✗	✗
Stackoverflow	<i>before-interaction</i>	✗	✗	✗
	<i>during-interaction</i>	✓	✗	✗
	<i>after-interaction</i>	✗	✗	✗
Wikipedia	<i>before-interaction</i>	✗	✗	✗
	<i>during-interaction</i>	✗	✗	✗
	<i>after-interaction</i>	✗	✗	✗
WordPress	<i>before-interaction</i>	✗	✓	✗
	<i>during-interaction</i>	✓	✗	✗
	<i>after-interaction</i>	✓	✗	✗
Overall	<i>before-interaction</i>	1	1	0
	<i>during-interaction</i>	4	1	2
	<i>after-interaction</i>	3	0	0

On the first criterion, the three PCSs with the highest number of usability problems were chosen, along with the three with the fewest. On the second criterion, five PCSs offered password policies, two provided password creation suggestions, and two presented password strength indicators. The combination of supporting features varied across the PCSs: only one PCS provided all three features, two provided two features, another two provided only one supporting feature. Moreover, one of the PCSs provided no help during the password creation process, so it was interesting to examine users' reaction to this instance of practice. Finally, on the third criterion the chosen sample varied with respect to the timing of these features' presentation. In seven instances, the features were presented in the *during-interaction* step, followed by three instances in the *after-interaction* and two instances in the *before-interaction* step.

For each PCS, a set of passwords was provided including valid and invalid passwords, and strong and weak passwords. These sets were the same as those used in the expert evaluation in Study 2 (see Section 4.2.3, Chapter 4).

5.2.4 Equipment and Materials

A MacBook Pro laptop running MacOS (v10.10) and Mozilla Firefox web browser (v33.1) were used. ScreenFlow software (v4.5) was employed to record the computer screen and the participants' think aloud protocols.

Participants were given a short demographic questionnaire (see Appendix B, Section B.3). This collected information about gender, age, native language, occupation, and daily usage of the internet and computers. Participants were also asked whether they had created an account in the last month for the PCSs in the evaluation.

5.2.5 Pilot of the Study Procedure

A pilot study was conducted with three PhD students from the Department of Computer Science to test the evaluation design and procedure. The study procedure was perceived as being smooth and interesting, and the task and instructions were clear to follow. However, several participants raised a concern about missing what was being presented on the screen as they were trying to copy the password from the instructions sheet and looking at the keyboard. Therefore, in the main study, the researcher attempted to solve this problem by asking participants to look at the password first and then to type in into the computer with the help of the researcher, who spelled out the password. The data from the pilot sessions were not included in the data analysis.

5.2.6 Procedure

Participants completed the evaluations in individual sessions lasting approximately 80 minutes, taking 15 minutes to evaluate each PCS. Sessions took place in the Interaction Labs in the Department of Computer Science at the University of York. The order of evaluating the PCSs was counterbalanced between the participants. For example, the

first participants evaluated the PCSs in the following order - PCS1, PCS2, PCS3, PCS4, PCS5, and PCS6, whereas the second participants evaluated the PCSs in the revised order. For each PCS, participants were provided with the same set of passwords used by the experts in Study 2 (see Section 4.2.3, Chapter 4) to create a password.

At the beginning of each session, the participant was briefed about the aim of study. The researcher emphasised that the study did not test the participant's password creation skills or involve using his or her own passwords, but instead comprised trying out a number of different PCSs with dummy passwords and commenting on how easy or difficult these systems were to use. Participants then read and signed an informed consent form (see Appendix A, Section A.2) and completed the demographic questionnaire. They also answered a question about their recent interaction with the PCSs to be evaluated in the last month. If the participant had created a password with any of the PCSs, he or she was excluded from evaluating that particular PCS.

Participants were instructed to think aloud while creating a password for each system, trying out the passwords provided and any others they wished to try. Once the participants were ready to start, the recording software was turned on, and the researcher started taking notes and gently prompting when appropriate.

For each PCS, whenever participants encountered a usability problem and commented on things they did not understand or things that confused them, they were asked to describe the problem and rate its severity using the same four-point scale (4 = catastrophic problem, 3 = major problem, 2 = minor problem, and 1 = cosmetic problem) as was used by the experts in Study 2. The participants then answered the following three questions about the PCS:

1. What do you think was the best thing about this PCS?
2. What do you think was the worst thing about this PCS?
3. What do you think most needs changing?

This procedure was repeated for all six PCS, as appropriate. Upon completion of the session, participants were debriefed and invited to ask any questions and give feedback on the study.

5.2.7 Data Analysis

A list was created of the usability problems identified by the participants during the evaluation sessions. A total of 753 instances of such problems were identified; 99 were discarded as not being usability problem (e.g. if a user was annoyed that a PCS did not allow him to have a weak password that he could remember well) or as being reported by only one participant. This left a total of 654 instances of usability problems (81 distinct usability problems) to include in the data analysis.

Similar to the expert evaluation, the usability problems were categorised using a content analysis following the three-step model, as discussed in Study 2 (see Section 4.2.7, Chapter 4). The same coding procedure was followed as in Study 2. Cohen's Kappa (K) (J. Cohen, 1960) was used to measure the agreement between coders, and the results showed an excellent level of agreement between coders on categories ($K=1$). For the three open-ended questions, a content analysis was also conducted, in which each response was assigned to a supporting feature for each PCS individually.

5.3 Results

This section presents the number of usability problems following the three-step model, followed by their categorisation. Subsequently, an analysis of three open-ended questions is conducted. Finally, the section compares the expert (Study 2, Chapter 4) and user (Study 3, present study) evaluations.

5.3.1 Usability Problems

In total, 654 instance of usability problems were encountered by at least two participants, with a mean of 4.56 per PCS (standard deviation = 0.51). Table 5.2 shows the number of instances of usability problem and distinct problems per PCS, as well as their mean severity rating. WordPress showed the highest percentage of instances

and distinct usability problems: 23.09% and 25.93%, respectively. It seems that the more interactivity there is, the more potential there is for usability problems to occur and vice versa. WordPress shows high level of interactivity by offering its statements of policy and creation suggestions at all three timings. Hence, the percentages of instances and distinct usability problems is high. On the other hand, Netflix and Wikipedia show low or no interactivity level and thus the percentages of instances and distinct usability problems is low. However, the mean of the severity rating of the usability problems in WordPress is not as severe as that of Netflix and Wikipedia. Thus, it seems that the number of usability problems does not predict the severity of the problems.

Table 5.2 Number of usability problem instances and distinct problems found by participants for each of the six PCSs, along with the mean severity ratings and range of problems per participant

PCS	No. of distinct problems (%)	No. of problem instances (%)	Mean severity rating (SD)	Mean problem instances/participant (SD)	Range/participant
Apple	16 (19.75)	126 (19.27)	2.52 (0.91)	5.25 (2.40)	1 – 10
DailyMail	11 (13.58)	129 (19.72)	2.83 (0.88)	5.38 (1.50)	3 – 9
Netflix	8 (9.88)	60 (9.17)	3.19 (0.83)	2.61 (0.94)	0 – 4
Stackoverflow	19 (23.46)	121 (18.50)	2.70 (1.00)	5.04 (1.79)	2 – 8
Wikipedia	6 (7.41)	67 (10.24)	3.31 (0.98)	2.79 (1.32)	0 – 5
WordPress	21 (25.93)	151 (23.09)	2.84 (0.87)	6.29 (1.65)	4 – 11
Total	81	654	2.90 (0.07)	4.56 (0.51)	-

Table 5.3 Number of distinct usability problems found by participants for each feature and step

Supporting features	Steps (timing of presentation)		
	<i>before-interaction</i>	<i>during-interaction</i>	<i>after-interaction</i>
	Frequency (%)	Frequency (%)	Frequency (%)
<i>Password policies</i>	6 (33.33)	15 (28.85)	0
<i>Password creation suggestions</i>	7 (38.89)	6 (11.54)	1 (9.09)
<i>Password strength indicators</i>	0	7 (13.46)	0
<i>Other feedback</i>	0	11 (21.15)	3 (27.27)
<i>Error messages</i>	0	7 (13.46)	4 (36.36)
<i>Other problems</i>	5 (27.78)	6 (11.54)	3 (27.27)
Total	18 (22.22)	52 (64.20)	11 (13.58)

Based on the three-step model, Table 5.3 indicates the number of usability problems found by participants in each step for each supporting feature. The results show that more than half of the usability problems (64.20%) occurred in the *during-interaction* step, when participants started entering their proposed password.

5.3.2 Categorisation of Usability Problems

Table 5.4 summarises the categorisation of usability problems. The same six main categories emerged as in the expert evaluation (Study 2, Chapter 4): *password policies*, *password creation suggestions*, *password strength indicators*, *other feedback*, *error messages*, and *other problems*. 32 subcategories were identified, 8 of which appeared in more than one main category. As shown in Table 5.4, *password policies*, *password creation suggestions*, and *error messages* categories were the most elaborated categories, with seven subcategories each, whereas *password strength indicators* category was the least elaborated with only two subcategories. The remaining two categories have between six and three subcategories each.

For the distinct problems, *password policies* category had the greatest number of problems (25.93%) while *password strength indicators* category had relatively small number of problems (8.64%). For the problem instances, *password policies* category had the greatest number of problems (24.31%) while *error messages* category had the fewest problem instances (10.70%). The mean severity ratings in the *error messages*, *other problems*, and *password strength indicators* categories were high. In terms of the number of PCSs, *password policies* and *other problems* categories had usability problems for all six evaluated PCSs. Out of the six categories, there were only two for which not all users identified usability problems.

Table 5.4 Number of instances of and distinct usability problems identified by users, with the number of PCSs in which they were encountered and mean severity ratings

Problem category	No. of distinct problems (<i>N</i> =81)	No. of problem instances (<i>N</i> = 654)	No. of PCSs (<i>N</i> =6)	No. of users (<i>N</i> =24)	Mean severity (SD)	Example
<i>Password policies</i>	21 (25.93%)	159 (24.31%)	6 (100%)	24 (100%)	2.67 (0.42)	
1.1. Not provided or provided too late	6	51	5	19	2.89 (0.92)	<i>There's no indication of what kind of password they want me to create. (P02)</i>
1.2. Information not detailed enough	2	13	2	12	3.00 (0.91)	<i>It doesn't tell you whether you need a capital letter or numbers – they specify just the length. (P21)</i>
1.3. Information too detailed	1	9	1	9	2.44 (1.01)	<i>It seems too specific and it's asking a lot of me. I'll probably read the first three and ignore the rest of it. (P17)</i>
1.4. Confusing/odd statement	8	54	2	22	2.57 (0.93)	<i>The first message says symbols or numbers and then when the password has a number they ask for symbols. They should be clear on that. (P09)</i>
1.5. Inconsistency of terms	1	19	1	19	2.50 (1.04)	<i>It's confusing changing the terms special characters to symbols. I would assume there are the same. (P01)</i>
1.6. Poor/unclear/confusing presentation	1	8	1	8	3.29 (0.76)	<i>The length definitely needs emphasis – I didn't notice that in the message. (P04)</i>
1.7. Querying the policy	2	5	2	5	2.00 (0.82)	<i>I think 4 is a bit small for a password and 60 characters is also an awful a lot in my opinion. It doesn't seem too safe! (P23)</i>
<i>Password creation suggestions</i>	14 (17.28%)	104 (15.90%)	5 (83.33%)	23 (95.83%)	2.60 (0.20)	
2.1. Not provided or provided too late	4	21	4	15	2.69 (0.96)	<i>I expect the system to help me to improve my password. (P14)</i>
2.2. Information not detailed enough	4	30	3	12	2.70 (0.72)	<i>It doesn't say I have to have a special character to make my password strong. (P03)</i>
2.3. Information too detailed	1	12	1	12	2.33 (0.52)	<i>That's good but don't know if anyone has enough time to read an essay on passwords. (P17)</i>

Problem category	No. of distinct problems (<i>N</i> =81)	No. of problem instances (<i>N</i> = 654)	No. of PCSs (<i>N</i> =6)	No. of users (<i>N</i> =24)	Mean severity (SD)	Example
2.4. Unclear how to create a good password	1	6	1	6	2.50 (1.00)	<i>My password strength is moderate but most of the dialog boxes are green, which makes me think 'what can I add to make it strong?' I would've hoped they'd give me some advice. (P12)</i>
2.5. Complex presentation	1	7	1	6	2.60 (1.14)	<i>The suggestions disappear – should be stated all the time (P05)</i>
2.6. Poor/unclear/confusing presentation	2	7	1	7	2.25 (0.96)	<i>I didn't notice that there is a link. (P11)</i>
2.7. Querying the creation suggestions	1	21	1	21	3.12 (0.86)	<i>That's ridiculous – no one would remember it. You should never ask a computer to generate your password. (P10)</i>
Password strength indicators	7 (8.64%)	93 (14.22%)	5 (83.33%)	24 (100%)	2.98 (0.20)	
3.1. Not provided	4	20	4	13	3.00 (1.12)	<i>Is it strong, is it weak? I do like to know the strength of my password. (P15)</i>
3.2. Querying the strength indicators	3	73	1	24	2.97 (0.84)	<i>I would expect the system to tell me that '123' is a common password, don't include it in your password, instead of only saying 'weak'. (P16)</i>
Other feedback	14 (17.28%)	126 (19.27%)	5 (83.33%)	24 (100%)	2.37 (0.40)	
4.1. Symbols not clear	2	18	2	15	1.86 (0.95)	<i>It seems it's accepted but what does the 'green tick' really mean? I think it means I completed that field. (P24)</i>
4.2. No feedback about valid/invalid password	4	45	4	21	2.90 (0.96)	<i>Their response when I do something OK is to stop telling me things are wrong, which I think a bit weak as it doesn't give you positive feedback. (P08)</i>
4.3. No feedback about matching/non-matching password	3	23	3	14	2.58 (1.02)	<i>It does not tell me if I got my confirmed password right. (P03)</i>
4.4. Feedback not accurate	2	16	2	15	2.57 (1.02)	<i>It said 'CAPS LOCK IS ON' even though it was not. (P22)</i>
4.5. Timing issue	1	2	1	2	1.00 (0.00)	<i>It's reacting a bit slowly. (P16)</i>

Problem category	No. of distinct problems (N=81)	No. of problem instances (N= 654)	No. of PCSs (N=6)	No. of users (N=24)	Mean severity (SD)	Example
4.6. Querying the security of the PCS	2	22	1	22	3.33 (0.66)	<i>The system contradicts itself by allowing me to use 'Password' but not 'password'. They apparently tell me in a previous message that it is the most common password on the web. (P24)</i>
Error messages	11 (13.58%)	70 (10.70%)	5 (83.33%)	23 (95.83%)	3.06 (0.32)	
5.1. Information not specific enough	3	28	2	19	2.71 (0.75)	<i>It's not saying what it wants me to add. It needs to be more specific. (P10)</i>
5.2. Poor/unclear/confusing presentation	1	9	1	9	3.11 (0.78)	<i>It gives me an error message and moderate strength. Is it going to let me use it? (P19)</i>
5.3. Inconsistent statement	2	13	1	12	2.62 (0.96)	<i>It's bad, it's saying the same thing in different ways. (P13)</i>
5.4. No visual distinction between error message and creation suggestions	1	6	1	6	3.40 (0.89)	<i>The error message in blue box will not draw my attention. (P04)</i>
5.5. Timing issue	1	3	1	3	3.33 (0.58)	<i>Why do you tell me 'password confirmation can't be empty'? I'm still in the password field but it draws my attention. (P15)</i>
5.6. Unclear/confusing language	2	9	1	9	2.25 (0.71)	<i>The word 'additional' in the error message is a bit vague. It doesn't specify which character I should add. (P17)</i>
5.7. Querying the security of the PCS	1	2	1	2	4.00 (0.00)	<i>They don't tell me that the password is too simple. (P23)</i>
Other problems	14 (17.28%)	102 (15.60%)	6 (100%)	24 (100%)	3.03 (0.22)	
6.1. Poor/unclear/confusing presentation	6	22	4	15	2.50 (1.06)	<i>The field is very small and the name of the field disappears when I type in it. (P22)</i>
6.2. Functionality lacking	3	33	3	22	3.00 (0.68)	<i>Not having a 'confirm password' – there's a big risk of making a typo and getting it wrong. (P18)</i>
6.3. Querying the security of the PCS	5	47	4	23	3.58 (0.66)	<i>It's a bit concerning that the system let me use '1234' as a password. (P01)</i>

5.3.3 Participant Perceptions of the Six PCSs

In addition to identifying usability problems, participants were also asked to answer three questions after evaluating each PCS. These questions aimed to gain better insight into their perceptions about the current practices of PCSs, and eventually to help in designing better PCSs. The first question aimed to gather information about the best practices that participants found helpful, whereas the second question asked about the worst practices that annoyed users, and the final question aimed to know what most needs changing in each PCSs.

On average participants reported 5.50 (standard deviation = 0.72) best practices, 5.63 (standard deviation = 0.72) worst practices, and 5.67 (standard deviation = 0.56) practices that needed changing in the PCSs. These responses were divided into six categories: *password policies*, *password creation suggestions*, *password strength indicators*, *other feedback*, *error messages*, and *other comments*.

5.3.3.1 Question 1: PCS Helpful Best Practices

As shown in Table 5.5, the distribution of the helpful best practices reported by participants across the six categories was varied, with most practices in the *other comments* and *other feedback* categories (44.02% and 25.16%, respectively). Very few best practices (5.03%) were reported regarding the *password strength indicators*, but this is because only two of the PCSs evaluated provided this supporting feature.

For the *other comments* category, the majority of responses concerned the overall impression of the design and process of password creation. Participants appreciated having a fairly straightforward and simple PCS like the Netflix and Wikipedia PCSs. Interestingly, these two PCSs are highly relaxed in terms of their password composition requirements. For example, Wikipedia has no password policy at all, so users can create passwords as simple (and weak) or as complex (and strong) as they like. In the *other feedback* category, participants found having it helpful to have instant feedback with coloured visual aids (such as ticks and exclamation marks). This was implemented in Apple, WordPress, and Netflix.

Table 5.5 Frequency of the helpful best practices across six categories, with the number of PCSs in which they occurred and the number of participants who reported these practices

Categories	No. of best practices (%) (N=159)	No. of PCSs (%) (N=6)	No. of participants (%) (N=24)	Example of responses
<i>Password policies</i>	17 (10.69)	5 (83.33)	12 (50.00)	<i>It was clear. They did say what it had to be in the password. (P09)</i>
<i>Password creation suggestions</i>	10 (6.29)	2 (33.33)	8 (33.33)	<i>I like the handy tips at the side – very clear and handy. (P12)</i>
<i>Password strength indicators</i>	8 (5.03)	2 (33.33)	7 (29.17)	<i>The best thing that they tell me is whether it's weak, or strong. (P24)</i>
<i>Other feedback</i>	40 (25.16)	6 (100)	21 (87.50)	<i>The best thing was having the tick boxes with the green light which show up as you fulfilled the criteria. (P08)</i>
<i>Error messages</i>	14 (8.81)	3 (50.00)	11 (45.83)	<i>The fact that it generates specific responses to what I just typed in. So, it's not a generic message applied to all the problems. (P22)</i>
<i>Other comments</i>	70 (44.02)	6 (100)	22 (45.83)	<i>Fairly straightforward and was not filling you with loads of information. (P01)</i>

5.3.3.2 Question 2: PCS Worst Practises

Most of the worst practices (40.49%) reported were related to the *other feedback* category. As shown in Table 5.6, they were encountered in all six PCSs by all 24 participants. Most of the worst practices in this category concerned either the feedback that participants received or the lack thereof. For example, participants felt insecure using PCSs that allowed a simple (and weak) password such as '1234', so they started to question the fact that this password was accepted. In addition, another poor practice of PCSs was not providing a confirmation on the validity of the password, as this would slow down participants' progress during the password creation process.

Very few worst practices (4.91%) were reported regarding the *password creation suggestions*. In fact, most were encountered in one PCS, DailyMail. For example, this PCS provides a strength indicator but it is not supported by creation suggestions to help participants know what to do to make their passwords stronger if they first try passwords which turn out to be weak. Participants wanted to have creation suggestions instead of having to guess how to improve their passwords.

Table 5.6 Frequency of the reported worst practices across the six categories, with the number of PCSs in which they were encountered and the number of participants who reported these practices

Categories	No. of worst practices (%) (N=163)	No. of PCSs (%) (N=6)	No. of participants (%) (N=24)	Example of responses
<i>Password policies</i>	29 (17.79)	6 (100)	14 (58.33)	<i>Lack of definite instruction. Everything seems to be a suggestion. (P18)</i>
<i>Password creation suggestions</i>	8 (4.91)	3 (50.00)	5 (20.83)	<i>Lack of clarity on what makes a good password. (P08)</i>
<i>Password strength indicators</i>	14 (8.59)	5 (83.33)	11 (45.83)	<i>The fact that it didn't tell me how secure my password was. (P22)</i>
<i>Other feedback</i>	66 (40.49)	6 (100)	24 (100)	<i>Doesn't give a confirmation that you've created the password correctly. (P21)</i>
<i>Error messages</i>	18 (11.04)	5 (83.33)	14 (58.33)	<i>When I get it really wrong it doesn't tell me what I've done wrong. (P03)</i>
<i>Other comments</i>	28 (17.18)	6 (100)	18 (75.00)	<i>I don't like the lack of 'confirm password' field. (P09)</i>

5.3.3.3 Question 3: What Needs to be Changed about the PCSs

Table 5.7 shows that the participants reported ways to improve all six PCSs across all but one category (the *error messages* category). Most of these improvements were related to the *password policies* (29.09%) and *other comments* (27.27%) categories. Conversely, very few changes (4.85%) were suggested in the *error messages* category. It appears that having a clear, helpful, and easy to understand PCSs is more important to overcome the issues that participants face when using a PCS than improving the presented error messages.

Regarding the improvements needed in the *password policies* category, participants thought that PCSs should present all the information related to the password composition (e.g. minimum/maximum characters required) beforehand and all at once, with clear language. They also thought about improving the complexity of the password policies. For example, if a PCS policy only concerned password length, the participants wanted this to be improved to include different character classes (i.e. uppercase letters, lowercase letters, digits, or symbols). Interestingly, even if the

passwords were not associated with high-risk accounts, for instance a Wikipedia account, which has no password policy at all, they thought this needed to be improved by adding some policies.

In terms of the *other comments* category, most of the changes participants requested/wanted were related to the legibility of the instructions, visibility of the input fields, and overall design of PCSs. Participants also thought that a ‘confirm passwords’ field needed to be added when it was missing, which is the case for Netflix and WordPress.

Table 5.7 Frequency of the reported characteristics that should be changed across the six categories, with the number of PCSs in which they were needed and number of participants who reported these characteristics

Categories	No. of changes (%) (N=165)	No. of PCSs (%) (N=6)	No. of participants (%) (N=24)	Example of responses
<i>Password policies</i>	48 (29.09)	6 (100)	20 (83.33)	<i>Password requirements should be told before you start entering your password. (P07)</i>
<i>Password creation suggestions</i>	18 (10.91)	6 (100)	12 (50.00)	<i>Probably when it says moderate, what can I add to make it strong? We need more advice and guidance. (P12)</i>
<i>Password strength indicators</i>	13 (7.88)	6 (100)	11 (45.83)	<i>Little bar telling you how strong your password is. (P18)</i>
<i>Other feedback</i>	33 (20.00)	6 (100)	19 (79.17)	<i>Stop sending mixed messages: I've got a really good password but it's too long. (P24)</i>
<i>Error messages</i>	8 (4.85)	4 (66.66)	6 (25.00)	<i>Tell me what I need to do to get my password right. (P03)</i>
<i>Other comments</i>	45 (27.27)	6 (100)	20 (83.33)	<i>The fact that it accepts numbers only – that needs changing the most. (P14)</i>

5.3.4 Comparison of Expert and User Evaluations

Since the experts evaluated a total of 12 PCSs and the users evaluated six, only a subset of the experts’ results was used for comparison with the users’ results. Thus, the comparison included the six PCSs that were assessed by both the experts and users: Apple, DailyMail, Netflix, Stackoverflow, Wikipedia, and WordPress. This meant that a total of 83 distinct usability problems from the experts and 81 distinct usability problems from the users were analysed. As discussed previously, each of these distinct

usability problems was assigned to a supporting feature and a step to which it belonged and in which it occurred. To match distinct usability problems between the users' and experts' list, three factors were considered: PCS, supporting feature, and step. Two problems were matched if they both belonged to the same PCS, were related to the same supporting feature, and occurred in the same step. The researcher matched the distinct problems separately and then together with her supervisor until there was complete agreement on the matching.

Table 5.8 Number of distinct usability problems found in expert and user evaluations for each of the six PCSs, along with the mean severity ratings

PCS	Experts only		Users only		Both experts and users		Overall
	No. of distinct problems	Mean (SD) severity rating	No. of distinct problems	Mean (SD) severity rating	No. of distinct problems	Mean (SD) severity rating	
Apple	10	1.78 (0.55)	3	2.79 (0.75)	13	2.29 (0.35)	26
DailyMail	14	2.08 (0.53)	5	2.83 (0.40)	6	2.69 (0.27)	25
Netflix	3	1.58 (0.52)	4	3.26 (0.87)	4	2.39 (0.71)	11
Stackoverflow	5	1.93 (0.49)	7	2.52 (0.54)	12	2.60 (0.46)	24
Wikipedia	4	2.58 (0.57)	2	3.89 (0.16)	4	2.68 (0.76)	10
WordPress	4	1.46 (0.42)	17	2.54 (0.70)	4	2.71 (0.50)	25
Total	40	1.90 (0.05)	38	2.97 (0.26)	43	2.45 (0.16)	121
	(33.06%)		(31.40%)		(35.54%)		

The two evaluations produced a pool of 121 distinct usability problems: 40 (33.06%) found by in expert evaluation only, 38 (31.40%) found in the user evaluation only, and 43 (35.54%) found in both expert and user evaluations. Table 5.8 presents the breakdown of the six PCSs, and shows the lack of an overall pattern of either expert or user evaluation in revealing a greater proportion of problems. In terms of severity ratings, the users were significantly more severe ($M = 2.97$; $Mdn = 2.83$) in their ratings than the experts ($M = 1.90$; $Mdn = 2.00$). A Mann-Whitney test confirmed this difference, $U = -4.70$, $p = .000$.

Table 5.9 summarises the categorisation of usability problems for the two evaluations. Six main categories emerged in both the expert and user evaluations: *password policies*, *password creation suggestions*, *password strength indicators*, *other feedback*, *error messages*, and *other problems*. Overall, nearly one-quarter of the usability problems were in the *password policies* category (30, 24.79%) while in

contrast, relatively few usability problems were found in the *password strength indicators* category (12, 9.91%). The other categories had relatively similar shares of usability problems: *password creation suggestions* had 20 (16.52%), *other feedback* 18 (14.87%), *error messages* 20 (16.52%), and *other problems* 21 (17.35%).

The experts encountered relatively more usability problems in the three key supporting features of the PCSs (i.e. *password policies*, *password creation suggestions*, and *password strength indicators*) and in the *error messages* category in comparison to the users. Conversely, the users experienced more problems with PCSs in terms of the *feedback* they provided and *other problems*, such as querying the security of the PCSs. In terms of the severity ratings, users gave higher ratings on the severity of the problems than experts across all six main categories.

Focussing on the three key supporting features, the results showed that both experts and users encountered usability problems related to the statements of policy in three subcategories: '*not provided or provided too late*', '*information not detailed enough*', and '*confusing/odd statement*'. Furthermore, both experts and users found similar usability problems for the statements of creation suggestions in only two subcategories (*not provided or provided too late*, '*information not detailed enough*'). For the strength indicators, both experts and users reported usability problems in regard to the provision of this feature '*Not provided*', but experts only encountered usability problems about the design aspects '*Poor colour contrast*'.

Table 5.9 Number of distinct usability problems identified by experts only, users only, and both experts and users divided into the main and subcategories of usability problems, along with the mean severity ratings

Problem category	Experts only		Users only		Both experts and users		Total no. of problems (N=121)
	No. of distinct problems	Mean (SD) severity rating	No. of distinct problems	Mean (SD) severity rating	No. of distinct problems	Mean (SD) severity rating	
Password policies	9 (30%)	1.73 (0.11)	7 (23.33%)	2.56 (0.17)	14 (46.66%)	2.35 (0.01)	30 (24.79%)
1.1. Not provided or provided too late	1	2.33	2	2.59 (0.48)	4	2.62 (0.57)	7
1.2. Information not detailed enough	3	2.14 (0.43)	1	2.25	1	2.77	5
1.3. Information too detailed	0	-	0	-	1	2.54	1
1.4. Confusing/odd statement	4	1.31 (0.28)	2	2.18 (0.72)	6	2.38 (0.56)	12
1.5. Inconsistency of terms	1	1.75	0	-	1	2.43	2
1.6. Poor/unclear/confusing presentation	0	-	1	3.29	0	-	1
1.7. Querying the policy	0	-	1	2.50	1	1.33	2
Password creation suggestions	6 (30%)	2.03 (0.04)	4 (20%)	2.67	10 (50%)	2.47 (0.30)	20 (16.52%)
2.1. Not provided or provided too late	2	2.13 (0.18)	1	3.50	3	2.33 (0.54)	6
2.2. Information not detailed enough	1	1.50	1	2.25	3	2.46 (0.12)	5
2.3. Information too detailed	0	-	0	-	1	2.25	1
2.4. Unclear how to create a good password	0	-	0	-	1	2.50	1
2.5. Unclear/confusing language	1	2.33	0	-	0	-	1
2.6. Complex presentation	2	2.17 (0.24)	0	-	1	2.25	3
2.7. Poor/unclear/confusing presentation	0	-	2	2.25 (0.35)	0	-	2
2.8. Querying the creation suggestions	0	-	0	-	1	3.00	1
Password strength indicators	5 (41.66%)	2.35	3 (25%)	3.13	4 (33.33%)	2.78 (0.05)	12 (9.91%)
3.1. Not provided	0	-	2	3.25 (1.06)	2	2.65 (0.03)	4
3.2. Poor colour contrast	3	2.06 (0.42)	0	-	0	-	3
3.3. Poor colour coding semantics	1	2.00	0	-	0	-	1
3.4. Querying the strength indicators	1	3.00	1	3.00	2	2.90 (0.10)	4

Problem category	Experts only		Users only		Both experts and users		Total no. of problems (N=121)
	No. of distinct problems	Mean (SD) severity rating	No. of distinct problems	Mean (SD) severity rating	No. of distinct problems	Mean (SD) severity rating	
Other feedback	4 (22.22%)	2.00	6 (33.33%)	2.32	8 (44.44%)	2.63 (0.22)	18 (14.87%)
4.1. Symbols not clear	1	2.00	1	1.67	1	1.64	3
4.2. No feedback about valid/invalid password	1	2.00	2	3.18 (0.45)	2	2.80 (0.51)	5
4.3. No feedback about match/non-matching password	0	-	1	2.75	2	2.77 (0.33)	3
4.4. Feedback not accurate	2	2.00 (0)	0	-	2	2.77 (0.33)	4
4.5. Timing issue	0	-	1	1.00	0	-	1
4.6. Querying the security of the PCS	0	-	1	3.00	1	3.26	2
Error messages	9 (45%)	2.38 (0.06)	6 (30%)	2.89 (0.26)	5 (25%)	2.71	20 (16.52%)
5.1. Information not specific enough	0	-	3	2.67 (0.33)	0	-	3
5.2. Poor/unclear/confusing presentation	2	2.21 (0.65)	0	-	1	3.00	3
5.3. Inconsistent statement	1	2.00	0	-	2	2.33 (0.05)	3
5.4. No visual distinction between error message and creation suggestion	1	3.00	0	-	1	3.00	2
5.5. Timing issue	4	2.42 (0.72)	0	-	1	2.50	5
5.6. Unclear/confusing language	0	-	2	2.00 (0.71)	0	-	2
5.7. Querying the security of the PCS	1	2.25	1	4.00	0	-	2
Other problems	7 (33.33%)	1.20 (0.24)	12 (57.14%)	3.05 (0.15)	2 (9.52%)	2.28	21 (17.35%)
6.1. Unclear/confusing language	2	1.13 (0.18)	0	-	0	-	2
6.2. Poor/unclear/confusing presentation	3	1.44 (0.51)	5	2.50 (0.59)	1	1.78	9
6.3. Poor colour contrast	1	1.25	0	-	0	-	1
6.4. Symbols not clear	1	1.00	0	-	0	-	1
6.5. Functionality lacking	0	-	2	3.21 (0.41)	1	2.78	3
6.6. Querying the security of the PCS	0	-	5	3.43 (0.29)	0	-	5

5.4 Discussion

The aims of this study were to understand the usability problems that users encounter in PCSs and to compare these problems with those identified in the expert evaluation (see Study 2, Chapter 4). This will provide insight into which of the two methods might be used to conduct usability evaluations of such systems. To address the first aim, six PCSs were evaluated with 24 potential users using a think-aloud protocol. A total of 654 instances and 81 distinct usability problems were identified. The number of usability problems found in the PCSs was surprisingly high: 109 problem instances and over 13 distinct problems per PCS, even though they are very small interactive systems.

The usability problems identified fell into six main categories and 32 subcategories. Similar to the expert evaluation (see Study 2, Chapter 4) the main categories were: the statement of *password policies*, the statement of *password creation suggestions*, *password strength indicators*, *feedback* to the user about the password creation process, *error messages* in case of password violations, and *other problems* that did not belong to any of the five other main categories.

Interestingly, the same set of patterns in the subcategories emerged from the user evaluation as from the expert evaluation, such as not providing enough information, providing too much information, and confusing statements and presentations. It seems that PCSs were criticised for providing too little instruction, yet also for providing too much instruction. Thus, there is a need to learn more about the forms and amount of instruction provided in PCSs to understand which are most helpful for users. To this end, Study 5 aimed to investigate different forms of instructions and to identify which are better for users. This is discussed in Chapter 7. In addition, it is worth noting that the same subcategory of problems appeared when users questioned the types of passwords which were allowed or not allowed by PCSs, as expert evaluators had. This may not strictly be a usability problem, but was included in the analysis because it

indicates further aspects of PCSs that may distract users, require them to put in cognitive effort, and in turn lead to poor password creation.

It is interesting to note that although only six PCSs were evaluated by users, the results from the user evaluation yielded similar findings to the expert evaluation (see Study 2, Chapter 4) with regard to the distribution of problems in the three-step model. More than half of usability problems occurred in the second step of the model, during the actual interaction with the system, when the users entered a password. Therefore, Study 6 aims to examine how this affects users when they create a password and how best to deploy the timing of presentation effectively to present these features to users, as will be discussed in Chapter 8.

The second aim of this study was to compare user and expert evaluation methods that might be used to conduct usability evaluations of PCSs. It was expected that in these relatively small systems, experts would be able to identify most of the usability problems. However, the overlap between the usability problems found by expert evaluation and user evaluation was 35.54%. This percentage is higher than what has been revealed in previous research on other types of interactive systems (Batra & Bishu, 2007; Hertzum et al., 2002; Jeffries, Miller, Wharton, & Uyeda, 1991; Petrie & Power, 2012). Interestingly, the results also showed that the users were more severe in their ratings than the experts were.

Given the experts' disappointing performance of missing 31.40% of the usability problems in such small systems, the categorisation of the problems was developed into a set of heuristics for evaluators and guidelines for developers specifically addressing PCSs which could guide both expert evaluation and the development of future PCSs. This is discussed in Study 8 in Chapter 10.

Overall, these findings may help researchers in usable security to learn more about the usability of PCSs, and help developers build usable PCSs. Having a usable PCS is important since it may lead users to create stronger passwords. However, these findings must be interpreted with caution for the following reasons. First, the users did

not use their own passwords during the evaluation, but were given a list of passwords instead. This might affect the interpretation of the results since the impact of PCSs on actual passwords was not considered. However, letting users use their own passwords in a lab environment could have affected the participation rate in the first place: people might not have been comfortable using their own passwords on real systems, or they could easily have used very common passwords, which might not have given them the chance to experience the PCSs' reactions to strong passwords. Second, users did not rate the perceived usability of the six PCSs using a standardised questionnaire. Given the number of PCSs that users had to evaluate in one session, the decision was made to ask users only three specific questions at the end of each PCS's evaluation to avoid participants dropping out due to the length of the session. The three questions also helped to gain better insight into users' perception of the current practices of PCS

5.5 Conclusions

All in all, the number of usability problems found by the users was high, in total 654 problem instances and 81 distinct problems, even though PCSs are very small interactive systems. It was therefore very important to see the impact of the usability problems and current practices in PCSs on users and their passwords to have a comprehensive understanding of current practices of PCSs. This is discussed in Study 4 in Chapter 6.

Chapter 6

The Effects of Current PCS Practices on Password Creation and Recall – *Study 4*

6.1 Introduction

While the user evaluation (see Study 3, Chapter 5) provided information about the usability problems users encountered with PCSs, the study presented in this chapter aims to help understand the impact of these problems and current practices in PCSs on users and the passwords they generate. Four PCSs were chosen from the set evaluated with users in Study 3 to design four types of mockup PCSs. An online study using MTurk was conducted to investigate how different practices of PCSs help or hinder users when they create and recall passwords. The following research questions were addressed in this study:

RQ 1. Are there differences in usability and password strength between different types of PCSs when users create passwords?

RQ 2. Are there differences in usability between different types of PCSs when users recall passwords?

6.2 Method

6.2.1 Design

This study used a between-participants design. The independent variable was the mockup of PCSs with four conditions: *Mockup1-Apple*, *Mockup2-DailyMail*, *Mockup3-Netflix*, and *Mockup4-WordPress*. The design of the different mockup PCSs

was based on the original design of four current PCSs that examined by users in Study 3 (see Chapter 5). The chosen PCSs were varied in terms of (1) number of usability problems, (2) whether and how they present the policy, (3) whether and how they present the suggestions, and (4) whether and how they present the strength indicators.

The study consisted of two parts: password creation (Part I) and password recall (Part II). In Part I, participants were asked to create a password using one of the PCSs, a role-playing approach was used to describe the password creation task. Participants were randomly assigned to one of the four conditions. In Part II, participants were asked to recall their password three days later.

Part I included two groups of dependent measures: those related to the usability of the PCSs, and those related to the strength of the passwords created by the participants. In terms of the usability of the PCSs, two types of measures were taken: efficiency and user satisfaction. Efficiency included three measures: (1) time to create and submit the password; (2) the number of keystrokes used for password creation entry; and (3) perceived workload to create a password using the NASA Task Load Index⁶ (NASA-TLX). User satisfaction included six measures: participants' ratings of (1) ease of use, (2) annoyingness, (3) helpfulness, (4) clarity, (5) amount of detail, and (6) their confidence in using the PCS. Participants were also offered the chance to explain their ratings of the PCS as an optional question.

In terms of password strength, two types of measure were taken: password characteristics and password guessability. Password characteristics included six measures: (1) password length, the total number of characters in the password, (2) the number of digits, (3) the number of uppercase letters, (4) the number of lowercase letters, (5) the number of symbols, and (6) the number of different character classes used in the password (i.e. digits, uppercase letters, lowercase letters, and symbols).

⁶ NASA Task Load Index: <http://humansystems.arc.nasa.gov/groups/tlx/>

Password guessability included one measure: the ability to guess the password by at least one of the five cracking approaches discussed in Ur et al. (2015). The set of passwords collected in this study was sent to a Password Guessability Service⁷ run by the Carnegie Mellon University Password Research Group. After several weeks of calculations, the service provided a guess number for each password uploaded under five cracking approaches: Hashcat, John the Ripper, Markov, Probabilistic Context-Free Grammar, and Minimum Across Automated Approaches.

For Part II, three dependent measures were taken of the usability of the PCSs: (1) the time to recall a password as a measure of efficiency, (2) the accuracy of recalling a password as a measure of effectiveness, and (3) participant's confidence in recalling the password correctly as a measure of user satisfaction.

6.2.2 Participants

257 people responded to the task in MTurk; however, 22 entries were excluded because their responses were incomplete. This left 235 participants included in the analysis. Participants were randomly assigned to one of the four conditions when they responded: this resulted in 59 participants producing data for *Mockup1-Apple*, 53 participants producing data for *Mockup2-DailyMail*, 60 participants producing data for *Mockup3-Netflix*, and 63 participants producing data for *Mockup4-WordPress*. Compensation was provided in the form of USD 0.70 (GBP 0.53) for completing Part I and a USD 0.70 bonus payment for returning and completing Part II.

Table 6.1 summarises the demographic characteristics of the participants in each condition and in total. Overall, 97 (41.30%) were females and 138 (58.70%) were males. The participants ranged in age from 20 to 67 years, with a mean age of 34.77 years (standard deviation = 10.57). A majority of participants (203, 86.40%) were

⁷ The Carnegie Mellon University Password Research Group's Password Guessability Service: <https://pgs.ece.cmu.edu>

native speakers of English, whereas the remaining had been speaking English for 21.88 years (standard deviation = 12.78). More than half of the participants (125, 53.20%) had a bachelor's degree. The level of education of the remaining participants ranged from postgraduate degree (50, 21.30%) to school qualification (15, 6.40%). In general, the majority of participants' major/career backgrounds were non-computing (154, 65.50%). On average, the majority of participants spent more than 6 hours a day online and using computers. As shown in Table 6.1, the participants' characteristics were similar between the four groups, except for the major/career background, for which the percentages in *Mockup2* group were divided evenly between computing and non-computing fields.

Table 6.1 Demographic characteristics (frequency and %) of participants in each group and overall

Characteristics		Groups of participant				Overall (N=235)
		Mockup1- Apple (N=59)	Mockup2- DailyMail (N=53)	Mockup3- Netflix (N=60)	Mockup4- WordPres s (N=63)	
Gender	<i>Female</i>	27 (45.80)	18 (34)	24 (40)	28 (44.40)	97 (41.30)
	<i>Male</i>	32 (54.20)	35 (66)	36 (60)	35 (55.60)	138 (58.70)
Language	<i>English</i>	50 (84.70)	45 (84.90)	56 (93.30)	52 (82.50)	203 (86.40)
	<i>Other</i>	9 (15.30)	8 (15.10)	4 (6.70)	11 (17.50)	32 (13.60)
Education	<i>School</i>	2 (3.40)	3 (5.70)	3 (5)	7 (11.10)	15 (6.40)
	<i>Diploma</i>	9 (15.30)	12 (22.60)	12 (20)	12 (19.10)	45 (17.90)
	<i>Bachelor's</i>	33 (55.90)	30 (56.60)	28 (46.70)	34 (54.0)	125 (53.20)
	<i>Master's</i>	14 (23.70)	7 (13.20)	13 (21.70)	10 (15.90)	44 (19.20)
	<i>Doctoral</i>	1 (1.70)	1 (1.90)	4 (6.70)	1 (1.60)	6 (2.60)
Major /Career	<i>Computing</i>	18 (30.50)	26 (49.10)	19 (31.70)	18 (28.60)	81 (34.50)
	<i>Non-computing</i>	41 (69.50)	27 (50.90)	41 (68.30)	45 (71.40)	154 (65.50)

6.2.3 Materials

Two web-based applications were developed to conduct the study. The first was the password creation application, which was used in Part I; and the second was the password recall application, which was used in Part II.

6.2.3.1 Password Creation Application

Figure 6.1 illustrates the structure of the password creation application. The application started with a homepage, on which the overall purpose of the study was explained, and an informed consent form was provided. Next, a scenario page was presented. To improve the ecological validity of the study and encourage participants to behave as they would normally do when creating a password, a role-playing approach was used to describe the password creation task (Fahl, Harbach, Acar, & Smith, 2013; Forget, Chiasson, van Oorschot, & Biddle, 2008; Just & Aspinall, 2009; Kelley et al., 2012; Komanduri et al., 2011; Schechter, Dhamija, Ozment, & Fischer, 2007).

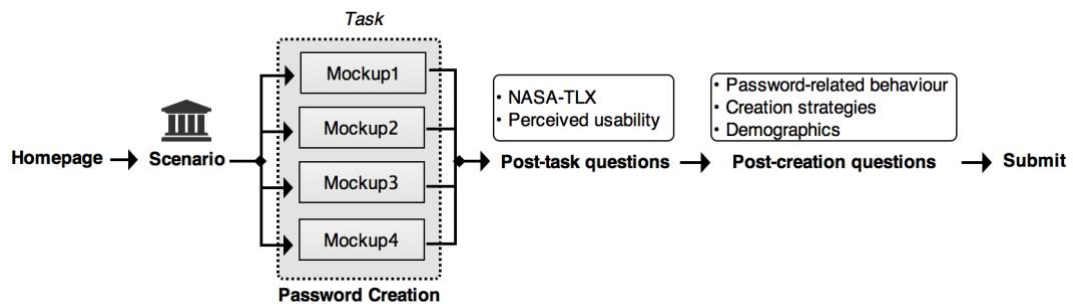


Figure 6.1 Structure of the password creation application used in the creation part

The scenario involved imagining a situation in which the participant's online bank account had been compromised and they needed to create a new password for their account using the PCS provided in the study. The scenario used in the study was the presented as follows:

Imagine that your online bank account has been attacked and has become compromised. You need to create a new password for your account, since your old password may now be known by the attackers. Because of the attack, your online bank system is also changing its password rules. The password you will now create should be easy to remember but hard for other people to guess. We will ask you to recall this password in three days so it is important that you remember it. Please take the steps you would normally take to remember your online bank password and protect this password as you normally would protect the password for your bank account. Please behave as you would if this were your real password!

The information provided in the scenario was based on one used by Kelley et al. (2012), but the context was different – they used an email account instead of bank account.

Four password creation pages were made (*Mockup1-Apple*, *Mockup2-DailyMail*, *Mockup3-Netflix*, and *Mockup4-WordPress*), one for each condition. When the participant successfully created a password (i.e. complying with the policy's requirement), the following acknowledgement message popped up: *'Well done! You have successfully created your password for your online bank account. When you are ready, click OK'*. After that, a post-task and a post-creation pages appeared, which asked participants to rate the PCS and to provide information about themselves.

6.2.3.1.1 Password Creation Page

Four PCSs were selected from the set evaluated by users in Study 3 (see Section 5.2.3 and Section 5.3.1, Chapter 5) to design the mockup PCSs used in the password creation pages. Table 6.2 lists the four mockup PCSs with the four chosen PCSs along with their supporting features and timing of presentation. The criteria for the chosen PCSs were (1) the number of usability problems identified by users, (2) the supporting features that these PCSs offered, and (3) the organisational structures of these features.

On the first criterion of usability problems, the two PCSs with the highest number of usability problems were chosen, and the two with the fewest usability problems. On the second criterion of supporting features, these four PCSs all offered password policies, two provided password creation suggestions, and two presented password strength indicators. The combination of supporting features varied across the four PCSs: only one PCS provided all three features, two provided two features, and one provided only one supporting feature. On the third criterion of organizational structure, the four PCSs varied with respect to the timing of presentation of the supporting features. For password policies, one PCS provided the policy at the during-interaction, two provided the policy at the during&after-interaction, and one PCS presented the policy at the before&after-interaction. Out of the two PCSs that provided

statements of creation suggestions, one PCSs presented the statement at the before-*interaction* and the other provided at the *during-*interaction**. The two PCSs which provided strength indicators, they presented this feature at the *during-*interaction**.

Table 6.2 The chosen sample for the four mockup PCSs

Mockup PCS	Chosen PCS	Supporting features & timing of presentation			No. of deployed usability problem instances	No. of deployed distinct usability problems
		<i>Password policies</i>	<i>Password creation suggestions</i>	<i>Password strength indicators</i>		
<i>Mockup1-Apple</i>	Apple	✓ during-&after- interaction	✗	✓ during- interaction	86/126 (68.25%)	13/16 (81.25%)
<i>Mockup2-DailyMail</i>	DailyMail	✓ during- interaction	✓ during- interaction	✓ during- interaction	71/129 (55.04%)	9/11 (81.82%)
<i>Mockup3-Netflix</i>	Netflix	✓ before-&after- interaction	✗	✗	39/60 (65.00%)	6/8 (75.00%)
<i>Mockup4-WordPress</i>	WordPress	✓ during-&after- interaction	✓ before- interaction	✗	100/151 (66.23%)	15/21 (71.43%)

For the purpose of this study, all the four mockup PCS conditions used the same policy statements, suggestions, and algorithm for calculating the strength of passwords. As a result of changing the policy requirements and the strength algorithm, not all usability problems found by users were deployed in the mockup PCSs.

The password policy used for all mockup PCSs in the study was that the password should contain at least 12 characters and include at least three of the four character classes (uppercase letters, lowercase letters, numbers, and symbols). This policy is recommended by Shay et al. (2014) as a usable and secure policy. Figures 6.2 to 6.5 show the presentation of the password policy in the four mockup PCSs. These varied as follows:

- *Mockup1-Apple* (see Figure 6.2) presented the policy as a dynamic built-point list at the *during-*interaction** step and as free text at the *after-*interaction** step;
- *Mockup2-DailyMail* (see Figure 6.3) presented the policy as free text at the *during-*interaction** step;

- *Mockup3-Netflix* (see Figure 6.4) presented the policy as free text at the *before-interaction* and *after-interaction* steps;
- *Mockup4-WordPress* (see Figure 6.5) presented the policy as free text at the *during-interaction* and *after-interaction* steps.

The password creation suggestions used in the two mockup PCSs was a mixture of different character classes. However, *Mockup4-WordPress* provided examples of symbols to use in the passwords. The presentation of the suggestion statements varied between the two mockup PCSs as follows:

- *Mockup2-DailyMail* (see Figure 6.3) presented the suggestion as free text at the *during-interaction* step;
- *Mockup4-WordPress* (see Figure 6.5) presented the suggestion as free text at the *before-interaction* step.

For the password strength indicator, the scoring algorithm used in this study was based on the equation proposed by Egelman et al. (2013): $strength = N \log_2 C$; where N is the total password length and C is the total character set size used (e.g. if digits and lowercase letters are used, these have individual character sets of 10 and 26 respectively, so the total character set size is 36). Following Egelman et al. (2013), the password strength was considered *weak* if the score was less than or equal to 56.53; *medium* if the score was greater than 56.53 and less than or equal to 71.09; and *strong* if the score was greater than 71.09. These numbers intervals were based on their pilot study with 51 participants. The presentation of the strength indicator was similar for the two mockup PCSs that used one:

- *Mockup1-Apple* (see Figure 6.2) presented the strength indicator as textual only underneath the password entry input field. The words displayed were *weak*, *moderate*, and *strong*; and
- *Mockup2-DailyMail* (see Figure 6.3) presented the strength indicator as textual only placed inside the password entry input field. The words displayed were *weak*, *medium*, and *strong*.

A final remark regarding is that two of the chosen PCSs, Apple and WordPress, prevented the use of common passwords. Therefore, a list of common passwords was prepared for this study to use in *Mockup1-Apple* and *Mockup4-WordPress*. The list was created from different sources on the web (SplashData and Keeper) that published the top 25 common passwords between 2011 and 2016.

(a) before-interaction step

(b) during-interaction step

(c) after-interaction step

Figure 6.2 Examples of design constructs provided on the *Mockup1-Apple* password creation page across the three timings of presentation

(a) before-interaction step

(b) during-interaction step

(c) after-interaction step

Figure 6.3 Examples of design constructs provided on the *Mockup2-DailyMail* password creation page across the three timings of presentation

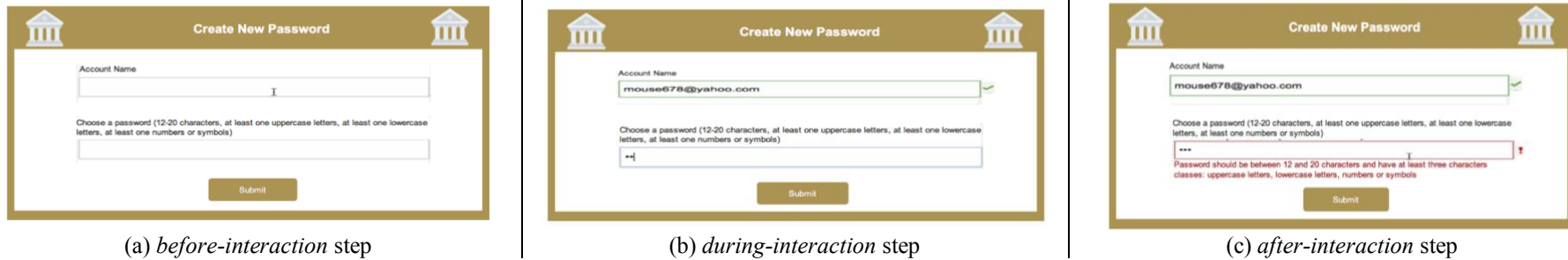


Figure 6.4 Examples of design constructs provided on the *Mockup3-Netflix* password creation page across the three timings of presentation

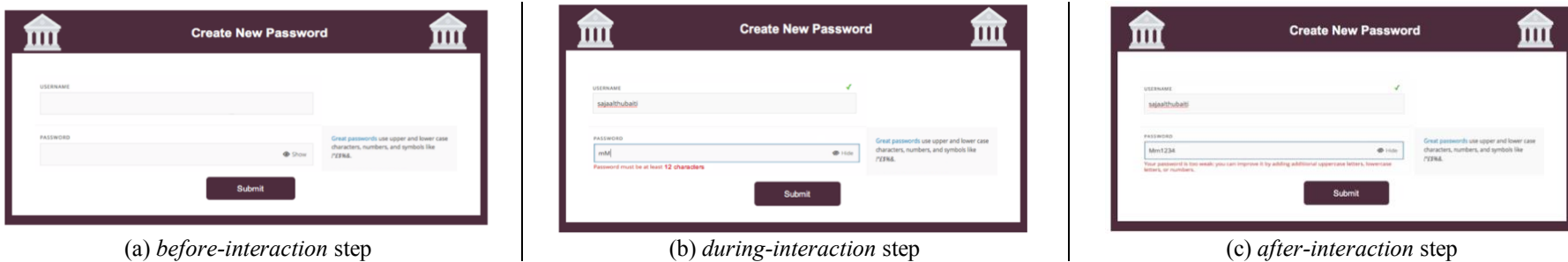


Figure 6.5 Examples of design constructs provided on the *Mockup4-WordPress* password creation page across the three timings of presentation

6.2.3.1.2 Post-Task Questions Page

A post-task questions page was provided at the end of the password creation task. The page contained questions that were split into the following two parts:

- Information about the perceived workload of creating a password, measured using the NASA-TLX. The NASA-TLX consists of six sub-scales: (1) mental demand, (2) physical demand, (3) temporal demand, (4) performance, (5) effort, and (6) participants' frustration level. The measures were rated using 20-point Likert items ranging from 1 (low) to 20 (high).
- Information about the perceived usability and satisfaction with the PCS. This part consisted of six measures: (1) ease of use, (2) annoyingness, (3) helpfulness, (4) clarity, (5) amount of detail, and (6) participants' confidence in using the PCS. These variables were measured using 5-point Likert items ranging from 1 (not at all easy/ extremely annoying/ not at all helpful/ not at all clear/ far too little detail/ not at all confident) to 5 (extremely easy to use/ not at all annoying/ extremely helpful/ extremely clear/ far too much detail/ extremely confident). There was also one optional open-ended question to give participants the chance to explain their ratings of the PCS.

6.2.3.1.3 Post-Creation Questions Page

Upon completion of the password creation application, a post-creation questions page was provided. The page contained questions that were split into three parts, as follows:

- Information about participants' password-related behaviours. This page asked about the participants' approximate number of online protected accounts and total number of passwords. It also included questions about participants' password-reuse and password-changing behaviours. In addition, there were questions about the participants' frequency of reading instructions when creating passwords and whether they had ever had a negative experience during the password creation process. This part also asked about participants' knowledge on creating a secure password. Finally, the participants were asked

about their familiarity with and usage of a password management system to store all passwords and remember only one password for that system.

- Information about the password creation strategies participants used to create their passwords in this study. This part consisted of two questions: the first asked about the methods participants used to create their passwords (e.g. based on a birthday, based on someone or something's name, or based on an address); and the second asked about the originality of the created passwords, and whether their created password was an entirely new, a reused, or modified one from a different account. This part was included to check whether participants put some effort in and tried their best in creating new passwords.
- Information about participants' demographic characteristics. This part contained questions about age, gender, native language, education, major/career, and finally, computer and internet usage.

6.2.3.2 Password Recall Application

Figure 6.6 shows an example of password recall page (*mockup4* recall page) provided in the recall application.

Try to remember the password you created three days ago using this password creation system and enter it below. The screenshot below is to help you remember the password creation system you used.

If you can't exactly remember your password, do not worry, just get as close as you can!

Create New Password

USERNAME

PASSWORD

Show

Great passwords use upper and lower case characters, numbers, and symbols like P@ssw0rd.

Submit

* 3. Please type in the password you created using this password creation system here

* 4. How confident are you that is the password?

Extremely confident Very confident Moderately confident Slightly confident Not at all confident

Next >>

Figure 6.6 A screenshot of the *mockup4* password recall page

For the password recall application, a recall page was developed for every password creation page (*Mockup1-Apple*, *Mockup2-DailyMail*, *Mockup3-Netflix*, and *Mockup4-WordPress*). Each recall page consisted of the task instructions, a screenshot of the mockup PCS used to create the password, a password entry field, and a question about participant's confidence in recalling the correct password. At the end of Part II, a post-recall questions page was presented asking about participants' methods for remembering their created passwords and their password management strategies.

6.2.4 Pilot of the Study Procedure

A pilot study was conducted with four PhD students from the Computer Science Department to test the study process and design. The study procedure was perceived as being smooth, and the instructions and given tasks were clear to follow. No issues were reported about the overall study procedure. The data from the pilot were not included in the data analysis.

6.2.5 Procedure

Participants were recruited via MTurk and directed to the password creation application. They were given a briefing about the study and an informed consent form at the beginning of the application. Participants were assured that the passwords would be anonymous and confidential, and would not be revealed. Participants confirmed their agreement and their understanding of the information provided in the briefing by clicking on the 'Next' button on the homepage. After that, participants were told to imagine that their online bank account had been compromised and that they needed to create a new password for their account. Participants were instructed to create a memorable but strong password. Each participant was assigned randomly to one of the four mockup PCS conditions: *Mockup1-Apple*, *Mockup2-DailyMail*, *Mockup3-Netflix*, and *Mockup4-WordPress*. After that, participants were asked to rate their perception of the workload required for the password creation task, as well as their satisfaction when creating the passwords. Upon completion of the password creation task, participants were asked to complete the post-creation questionnaire.

Three days after finishing Part I, participants were invited via email through MTurk to complete Part II. They had to respond within 24 hours of receiving the invitation. Participants were asked to recall their passwords and to complete the post-recall questionnaire.

6.2.6 Data Analysis

Kolmogorov-Smirnov and Shapiro-Wilk tests were used to test for normality on all dependent measures. All dependent measures were significantly non-normal ($p < 0.05$), for both tests. Therefore, non-parametric statistics were used throughout the analysis. Kruskal-Wallis tests (H statistic) were used to assess the significance of differences in between the PCS conditions. Furthermore, when the dependent measures were of a frequency type, chi-square (χ^2 statistics) tests were used to measure the association among the categories (i.e. number of character classes, policy compliance, suggestion compliance, password guessability, and accuracy).

During the data preparation process, outliers were identified and adjusted for the following dependent measures: creation time, keystrokes, password length, number of digits, number of uppercase letters, number of lowercase letters, number of symbols, and recall time. The method used to adjust the outliers was data trimming, in which outliers were replaced by boundary values. The boundary values were calculated using the median and semi-interquartile range (SIQR). The formula for replacing the high outliers was $Median + SIQR$; the formula for replacing the low outliers was $Median - SIQR$.

6.3 Results

The results of this study are divided into two sections: first, the password creation results from Part I, and second, the password recall results from Part II.

6.3.1 Password Creation

The following section presents the results regarding the usability of the four mockup PCS conditions and the strength of the passwords created in these conditions. The four

mockup PCS conditions varied in different factors: (1) number of usability problems, (2) whether and how they present the policy, (3) whether and how they present the suggestions, and (4) whether and how they present the strength indicators. In this manner, comparing the four mockup PCS conditions would not inform us about which factor is causing the effect, so the decision was made to conduct the analysis on each factor separately. The results of the first (effect of usability problems) and second (effect of policy presentation) factors are presented next. Since the third and fourth factors produced the same results of the second factor, their results are not reported for space consideration.

For the first factor, the four mockup PCSs were divided into four levels based on the number of usability problems they have. The rank of the four mockup PCSs were exactly the same for both the number of distinct problems and problems instances. Therefore, four conditions were examined in terms of the level of usability problems: very low (*Mockup3-Netflix*), low (*Mockup2-DailyMail*), high (*Mockup1-Apple*), and very high (*Mockup4-WordPress*). For the second factor, the four mockup PCSs were divided into three levels based on the policy timing of presentation. Therefore, three conditions were examined in terms of the policy presentation: during-interaction (*Mockup2-DailyMail*), before&after-interaction (*Mockup3-Netflix*), and during&after-interaction (*Mockup1-Apple* and *Mockup4-WordPress*).

6.3.1.1 Usability of the PCSs

6.3.1.1.1 Efficiency Measures

- *Effect of number of usability problems*

Table 6.3 summarises the results for the three efficiency measures in terms of number of usability problems.

Creation time. There was a significant difference in the creation time between the four conditions of usability problems ($H(3) = 18.74, p < .001$). Participants in the very high (*Mockup4-WordPress*) condition spent significantly less time creating passwords than those in the other three conditions. The pairwise comparisons confirmed this difference, as shown in Table 6.4.

Table 6.3 Mean (median) creation time, keystrokes, and perceived workload measures between the usability problems conditions

	Usability problems conditions				<i>p</i> value	
	Very low (Mockup3- Netflix)	Low (Mockup2- DailyMail)	High (Mockup1- Apple)	Very high (Mockup4- WordPress)		
Creation time	60.06 (52.00)	72.33 (57.00)	69.43 (57.00)	44.96 (32.00)	.000	
Keystroke	62.65 (58.00)	60.33 (60.00)	71.67 (58.00)	56.27 (50.00)	n.s.	
Perceived workload	Mental demand	10.93 (11.50)	11.92 (13.00)	11.53 (13.00)	11.48 (12.00)	n.s.
	Physical demand	6.22 (3.00)	8.15 (6.00)	6.07 (4.00)	6.62 (4.00)	n.s.
	Temporal demand	8.37 (7.00)	11.09 (12.00)	8.49 (9.00)	9.00 (10.00)	n.s.
	Performance	13.70 (17.00)	13.32 (17.00)	13.46 (17.00)	14.16 (18.00)	n.s.
	Effort	12.23 (13.50)	11.94 (12.00)	10.83 (11.00)	10.57 (11.00)	n.s.
	Frustration	8.27 (6.00)	9.30 (10.00)	7.59 (6.00)	8.40 (8.00)	n.s.
	<i>Overall</i>	9.95 (10.00)	10.96 (11.00)	9.66 (9.83)	10.04 (10.50)	n.s.

Keystrokes. There was no significant difference in the number of keystrokes used to create a password between the usability problems conditions ($H(3) = 2.32$, $p = .508$).

Perceived workload. There was no significant difference in the overall perceived workload as measured by the NASA-TLX to create a password between the four usability problems conditions ($H(3) = 2.80$, $p = .424$). There were also no significant differences in any of the individual NASA-TLX scales (Mental Demand: $H(3) = 2.80$, $p = .845$; Physical Demand: $H(3) = 2.80$, $p = .405$; Temporal Demand: $H(3) = 2.80$, $p = .087$; Performance: $H(3) = 2.80$, $p = .487$; Effort: $H(3) = 2.80$, $p = .274$; Frustration: $H(3) = 2.80$, $p = .673$).

Table 6.4 Pairwise comparisons of creation time measure between the usability problems conditions

		Very low (Mockup3- Netflix)	Low (Mockup2- DailyMail)	High (Mockup1- Apple)	Very high (Mockup4- WordPress)
Creation time	Very low (Mockup3-Netflix)	-	15.26	17.64	29.79*
	Low (Mockup2-DailyMail)		-	2.38	45.05*
	High (Mockup1-Apple)			-	47.43*
	Very high (Mockup4-WordPress)				-

Note. * denotes a significant pairwise comparison, $p < .05$.

- **Effect of policy presentation**

Table 6.5 summarises the results for the three efficiency measures in terms of policy presentation.

Table 6.5 Mean (median) creation time, keystrokes, and perceived workload measures between different policy presentation conditions

	Policy presentation conditions			<i>p</i> value
	before&after- interaction (<i>Mockup3-Netflix</i>)	during- interaction (<i>Mockup2-DailyMail</i>)	during&after- interaction (<i>Mockup1-Apple</i> and <i>Mockup4-Wordpress</i>)	
Creation time	60.06 (52.00)	72.33 (57.00)	56.80 (42.50)	n.s.
Keystroke	62.65 (58.00)	60.33 (60.00)	63.72 (52.00)	n.s.
Perceived workload	Mental demand	10.93 (11.50)	11.92 (13.00)	n.s.
	Physical demand	6.22 (3.00)	8.15 (6.00)	n.s.
	Temporal demand	8.37 (7.00)	11.09 (12.00)	.040
	Performance	13.70 (17.00)	13.32 (17.00)	n.s.
	Effort	12.23 (13.50)	11.94 (12.00)	n.s.
	Frustration	8.27 (6.00)	9.30 (10.00)	n.s.
	<i>Overall</i>	9.95 (10.00)	10.96 (11.00)	9.86 (10.17)

Creation time. There was no significant difference in the creation time between the three types of policy presentation ($H(2) = 3.91$, $p = .141$).

Keystrokes. There was no significant difference in the number of keystrokes used to create a password between the types of policy presentation ($H(2) = 0.39$, $p = .824$).

Perceived workload. There was no significant difference in the overall perceived workload as measured by the NASA-TLX to create a password between the three types of policy presentation ($H(2) = 2.49$, $p = .288$). There were also no significant differences in five of the individual NASA-TLX scales (Mental Demand: $H(2) = 0.81$, $p = .666$; Physical Demand: $H(2) = 2.86$, $p = .239$; Performance: $H(2) = 1.42$, $p = .492$; Effort: $H(2) = 3.86$, $p = .145$; Frustration: $H(2) = 1.25$, $p = .536$). However, a significant difference was found in the ratings of temporal demand to create passwords between the three types of policy presentation ($H(2) = 6.42$, $p = .040$). Participants in the during-interaction (*Mockup2-DailyMail*) condition rated the temporal demand required to create passwords significantly higher than those in the other two conditions. The pairwise comparisons confirmed this difference (see Table 6.6).

Table 6.6 Pairwise comparisons of ratings of temporal demand measure between policy presentation conditions

		before&after- interaction (Mockup3- Netflix)	during- interaction (Mockup2- DailyMail)	during&after- interaction (Mockup1-Apple and Mockup4-Wordpress)
Temporal demand	before&after-interaction (Mockup3-Netflix)	-	-28.87*	-3.45
	during-interaction (Mockup2-DailyMail)		-	25.42*
	during&after-interaction (Mockup1-Apple and Mockup4- Wordpress)			-

Note. * denotes a significant pairwise comparison, $p < .05$.

6.3.1.1.2 User Satisfaction Measures

- *Effect of number of usability problems*

Table 6.7 summarises the results for the six user satisfaction measures in terms of the number of usability problems. There was no significant difference in the ratings of ease of use ($H(3) = 1.51$, $p = .681$), annoyingness ($H(3) = 1.06$, $p = .787$), helpfulness ($H(3) = 0.78$, $p = .853$), clarity ($H(3) = 3.59$, $p = .309$), amount of detail ($H(3) = 0.67$, $p = .882$), or confidence ($H(3) = 2.98$, $p = .395$) between the four usability problems conditions.

Table 6.7 Mean (median) ratings of ease of use, annoyingness, helpfulness, clarity, amount of details, and participants' confidence measures between the usability problems conditions

	Usability problems conditions				p value
	Very low (Mockup3- Netflix)	Low (Mockup2- DailyMail)	High (Mockup1- Apple)	Very high (Mockup4- WordPress)	
Ease of use	2.43 (2.00)	2.72 (3.00)	2.51 (2.00)	2.60 (2.00)	n.s.
Annoyingness	2.72 (3.00)	2.87 (2.00)	2.61 (2.00)	2.79 (3.00)	n.s.
Helpfulness	3.82 (4.00)	3.06 (3.00)	3.20 (3.00)	3.98 (4.00)	n.s.
Clarity	3.82 (4.00)	3.74 (4.00)	4.08 (3.00)	3.98 (4.00)	n.s.
Amount of detail	2.17 (2.00)	2.19 (2.00)	2.22 (2.00)	2.11 (2.00)	n.s.
Confidence	3.35 (3.00)	3.45 (4.00)	3.96 (4.00)	3.59 (4.00)	n.s.

- *Effect of policy presentation*

Table 6.8 summarises the results for the six user satisfaction measures in terms of policy presentation. There was no significant difference in the ratings of ease of use ($H(2) = 1.50, p = .473$), annoyingness ($H(2) = 0.51, p = .775$), helpfulness ($H(2) = 0.78, p = .679$), clarity ($H(2) = 3.37, p = .186$), amount of detail ($H(2) = 0.01, p = .998$), or confidence ($H(2) = 2.78, p = .252$) between the three policy presentation conditions.

Table 6.8 Mean (median) ratings of ease of use, annoyingness, helpfulness, clarity, amount of details, and participants' confidence measures between the policy presentation conditions

	Policy presentation conditions			<i>p</i> value
	before&after- interaction (<i>Mockup3-Netflix</i>)	during- interaction (<i>Mockup2-DailyMail</i>)	during&after- interaction (<i>Mockup1-Apple</i> and <i>Mockup4-Wordpress</i>)	
Ease of use	2.43 (2.00)	2.72 (3.00)	2.56 (2.00)	n.s.
Annoyingness	2.72 (3.00)	2.87 (2.00)	2.70 (2.50)	n.s.
Helpfulness	3.82 (4.00)	3.06 (3.00)	3.16 (3.00)	n.s.
Clarity	3.82 (4.00)	3.74 (4.00)	4.03 (4.00)	n.s.
Amount of detail	2.17 (2.00)	2.19 (2.00)	2.16 (2.00)	n.s.
Confidence	3.35 (3.00)	3.45 (4.00)	3.64 (4.00)	n.s.

In summary, the usability did not differ significantly among the four mockup PCSs, except the creation time and perceived temporal demand. Unexpectedly, participants spent significantly less time creating passwords with the *Mockup4-WordPress* PCS which had the highest number of usability problems. Furthermore, the level of temporal demand was perceived to be significantly the highest in the *Mockup2-DailyMail* PCS which provided the policy statement during password entry.

6.3.1.2 Strength of Password

6.3.1.2.1 Password Characteristics

- *Effect of number of usability problems*

Table 6.9 summarises the results for the password characteristic measures in terms of the usability problems; Table 6.10 presents the results of the pairwise comparison for the number digits, number uppercase letters, and number of symbols measures between the usability problems conditions.

Table 6.9 Mean (median) ratings of the password length, number of digits, number of uppercase letters, number of lowercase letters, and number of symbols measures between usability problems conditions

	Usability problems conditions				<i>p</i> value
	Very low (<i>Mockup3-Netflix</i>)	Low (<i>Mockup2-DailyMail</i>)	High (<i>Mockup1-Apple</i>)	Very high (<i>Mockup4-WordPress</i>)	
Password length	13.68 (13.50)	13.78 (13.00)	13.36 (13.00)	14.02 (13.00)	n.s.
Number of digits	3.08 (3.00)	3.12 (3.50)	3.81 (4.00)	3.62 (4.00)	.041
Number of uppercase	1.34 (1.00)	1.33 (1.00)	1.11 (1.00)	0.91 (1.00)	.000
Number of lowercase	7.90 (8.00)	7.30 (8.00)	7.51 (7.00)	7.06 (7.00)	n.s.
Number of symbols	0.94 (1.00)	0.90 (1.00)	0.50 (0.00)	1.38 (1.00)	.000

Password length. There was no significant difference in the length of the passwords created between the usability problems conditions ($H(3) = 1.27, p = .736$).

Number of digits. There was a significant difference in the number of digits used in passwords between the four usability problems conditions ($H(3) = 8.24, p = .041$). Passwords created in the high number of usability problems condition (*Mockup1-Apple*) included significantly more digits than those created in the very low (*Mockup3-Netflix*) and low (*Mockup2-DailyMail*) conditions. The pairwise comparison confirmed this difference (see Table 6.10).

Number of uppercase letters. There was a significant difference in the number of uppercase letters used in passwords between the four usability problems conditions ($H(3) = 20.85, p < .001$). Passwords created with very low (*Mockup3-Netflix*) and low (*Mockup2-DailyMail*) number of usability problems conditions included significantly more uppercase letters than those created with high (*Mockup1-Apple*) and very high (*Mockup4-WordPress*) conditions. However, there was no significant difference in the number of uppercase letters used in passwords between very low (*Mockup3-Netflix*) and low (*Mockup2-DailyMail*) conditions. The pairwise comparison confirmed these differences (see Table 6.10).

Number of symbols. There was a significant difference in the number of symbols used in passwords between the four usability problems conditions ($H(3) = 27.31, p < .001$). Passwords created in the very high (*Mockup4-WordPress*) number of usability

problems condition contained significantly more symbols than those created in other three conditions. On the other hand, passwords created in the high (*Mockup1-Apple*) number of usability problems condition contained significantly fewer symbols than those created in the other three conditions. These differences were confirmed in the pairwise comparison (see Table 6.10).

Table 6.10 Pairwise comparison for the number of digits, number of uppercase letters, and number of symbols measures between usability problems conditions

		Very low (<i>Mockup3-Netflix</i>)	Low (<i>Mockup2-DailyMail</i>)	High (<i>Mockup1-Apple</i>)	Very high (<i>Mockup4-WordPress</i>)
Number of digits	Very low (<i>Mockup3-Netflix</i>)	-	6.24	31.22*	-22.50
	Low (<i>Mockup2-DailyMail</i>)		-	24.98*	-16.26
	High (<i>Mockup1-Apple</i>)			-	8.72
	Very high (<i>Mockup4-WordPress</i>)				-
Number of uppercase	Very low (<i>Mockup3-Netflix</i>)	-	-3.21	-23.27*	40.14*
	Low (<i>Mockup2-DailyMail</i>)		-	-20.05	36.92*
	High (<i>Mockup1-Apple</i>)			-	16.87
	Very high (<i>Mockup4-WordPress</i>)				-
Number of symbols	Very low (<i>Mockup3-Netflix</i>)	-	3.62	-31.34*	-29.20*
	Low (<i>Mockup2-DailyMail</i>)		-	-34.97*	-29.20*
	High (<i>Mockup1-Apple</i>)			-	-60.54*
	Very high (<i>Mockup4-WordPress</i>)				-

Note. * denotes a significant pairwise comparison, $p < .05$.

Number of lowercase letters. There was no significant difference in the number of lowercase letters used in passwords between the four usability problems conditions ($F(3) = 3.09$, $p = .378$).

Number of password character classes. All participants had to comply with a policy of at least three character classes. Thus, this dependent variable examined whether the created passwords significantly included three or four classes. The results showed that there was no significant difference in all four usability problems conditions: very low (*Mockup3-Netflix*: $\chi^2(1) = 0.27$, $p = .606$), low (*Mockup2-DailyMail*: $\chi^2(1) = 0.02$, $p = .891$), high (*Mockup1-Apple*: $\chi^2(1) = 0.83$, $p = .362$), and very high (*Mockup4-WordPress*: $\chi^2(1) = 0.78$, $p = .378$). The distribution of the password character classes across the different usability problems conditions is shown in Figure 6.7.

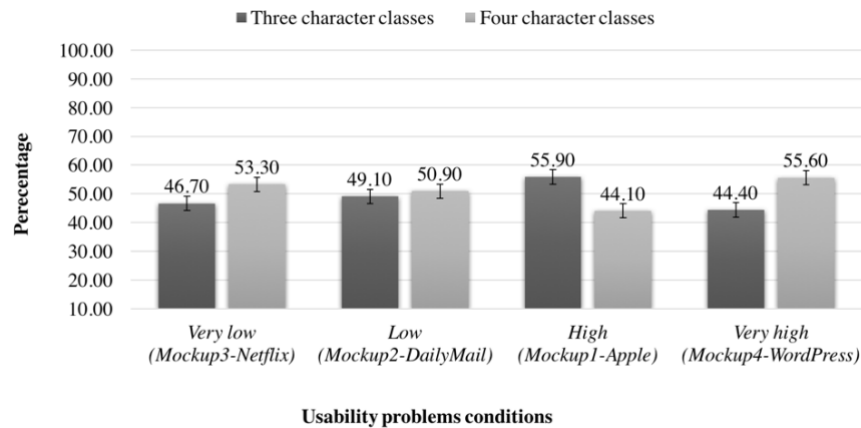


Figure 6.7 Percentage of password character classes across the four usability problems conditions

- ***Effect of policy presentation***

Table 6.11 summarises the results for the password characteristic measures in terms of the policy presentation; Table 6.12 presents the results of the pairwise comparison for the number digits and number uppercase letters measures between the policy presentation conditions.

Password length. There was no significant difference in the length of the passwords created between the three policy presentation conditions ($H(2) = 0.37$, $p = .830$).

Table 6.11 Mean (median) ratings of the password length, number of digits, number of uppercase letters, number of lowercase letters, and number of symbols measures between policy presentation conditions

	Policy presentation conditions			<i>p</i> value
	before&after-interaction (Mockup3-Netflix)	during-interaction (Mockup2-DailyMail)	during&after-interaction (Mockup1-Apple and Mockup4-Wordpress)	
Password length	13.68 (13.50)	13.78 (13.00)	13.70 (13.00)	n.s.
Number of digits	3.08 (3.00)	3.12 (3.50)	3.71 (4.00)	.021
Number of uppercase	1.34 (1.00)	1.33 (1.00)	1.01 (1.00)	.000
Number of lowercase	7.90 (8.00)	7.30 (8.00)	7.28 (7.00)	n.s.
Number of symbols	0.94 (1.00)	0.90 (1.00)	0.95 (1.00)	n.s.

Number of digits. There was a significant difference in the number of digits used in passwords between the three policy presentation conditions ($H(2) = 7.72$, $p = .021$). Passwords created in the during&after interaction condition (*Mockup1-Apple* and

Mockup4-WordPress) included significantly more digits than the before&after-interaction (*Mockup3-Netflix*) condition. The pairwise comparison confirmed this difference (see Table 6.12).

Table 6.12 Pairwise comparison for the number of digits, number of uppercase letters, and number of symbols measures between policy presentation conditions

		before&after- interaction <i>(Mockup3- Netflix)</i>	during- interaction <i>(Mockup2- DailyMail)</i>	during&after- interaction <i>(Mockup1-Apple and Mockup4-WordPress)</i>
Number of digits	before&after-interaction <i>(Mockup3-Netflix)</i>	-	-6.24	-26.71*
	during-interaction <i>(Mockup2-DailyMail)</i>		-	-20.48
	during&after-interaction <i>(Mockup1-Apple and Mockup4- WordPress)</i>			-
Number of uppercase	before&after-interaction <i>(Mockup3-Netflix)</i>	-	3.21	31.98*
	during-interaction <i>(Mockup2-DailyMail)</i>		-	28.77*
	during&after-interaction <i>(Mockup1-Apple and Mockup4- WordPress)</i>			-

Note. * denotes a significant pairwise comparison, $p < .05$.

Number of uppercase letters. There was a significant difference in the number of uppercase letters used in passwords between the three policy presentation conditions ($H(2) = 18.00, p < .001$). Passwords created in the before&after-interaction (*Mockup3-Netflix*) and during-interaction (*Mockup2-DailyMail*) conditions included significantly more uppercase letters than those created in the during&after-interaction (*Mockup4-WordPress* and *Mockup1-Apple*) condition. However, there was no significant difference in the number of uppercase letters used in passwords between before&after-interaction (*Mockup3-Netflix*) and during-interaction (*Mockup2-DailyMail*) conditions. The pairwise comparison confirmed these differences (see Table 6.12).

Number of lowercase letters. There was no significant difference in the number of lowercase letters used in passwords between the policy presentation conditions ($H(2) = 2.51, p = .285$).

Number of symbols. There was no significant difference in the number of symbols used in passwords between the policy presentation conditions ($H(2) = 0.14, p = .935$).

Number of password character classes. As discussed previously, all participants had to comply with a policy of at least three character classes. Thus, this dependent variable examined whether the created passwords significantly included three or four classes. The results showed that there was no significant difference in all three types of policy conditions: before&after-interaction (*Mockup3-Netflix*: $x^2(1) = 0.27, p = .606$), during-interaction (*Mockup2-DailyMail*: $x^2(1) = 0.02, p = .891$), and during&after-interaction (*Mockup1-Apple* and *Mockup4-WordPress*: $x^2(1) = 0.00, p = 1.00$). The distribution of the password character classes across the different policy presentation conditions is shown in Figure 6.8.

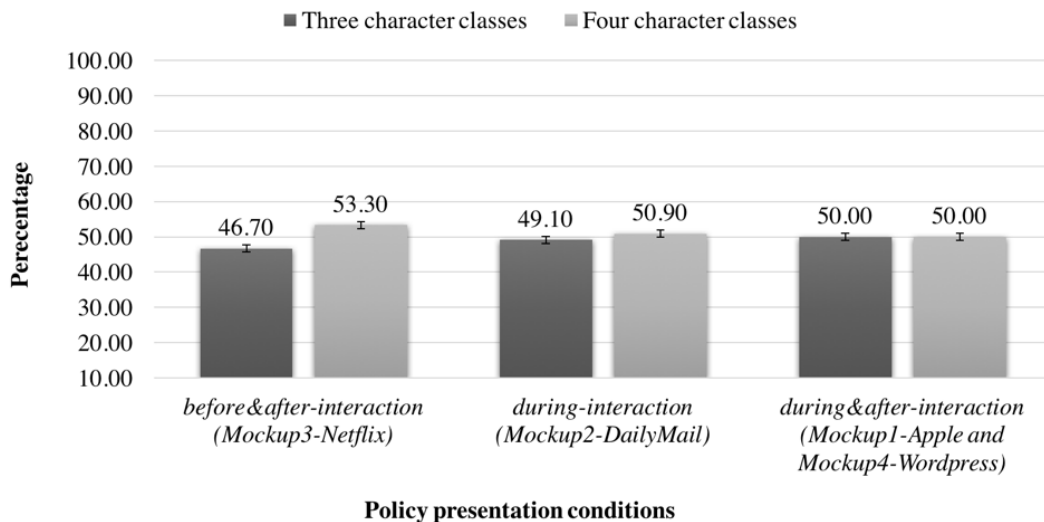


Figure 6.8 Percentage of password character classes across the three policy presentation conditions

6.3.1.2.2 Password Guessability

The ability to guess the created passwords using five cracking approaches (based on Ur et al., 2015) was also used to measure password strength. Overall, more than half of the created passwords (156, 66.38%) were not guessed by any of the five cracking approaches. In contrast, a third of them (79, 33.62%) were guessed by at least one of the approaches.

- *Effect of number of usability problems*

Figure 6.9 shows the percentage of password guessability for the four usability problems conditions. Looking at the four conditions individually, the results showed that there was a significant difference in three conditions: very low (*Mockup3-Netflix*: $x^2(1) = 8.07$, $p = .005$), low (*Mockup2-DailyMail*: $x^2(1) = 9.98$, $p = .002$), and very high (*Mockup4-WordPress*: $x^2(1) = 8.40$, $p = .004$), but not in the high (*Mockup1-Apple*: $x^2(1) = 1.37$, $p = .241$). The condition with low number of usability problems (*Mockup2-DailyMail*) had the highest percentage of passwords that were not guessable (71.70%), whereas the conditions with very highest (*Mockup4-WordPress*) number of problems had the highest percentage of guessable ones (31.75%).

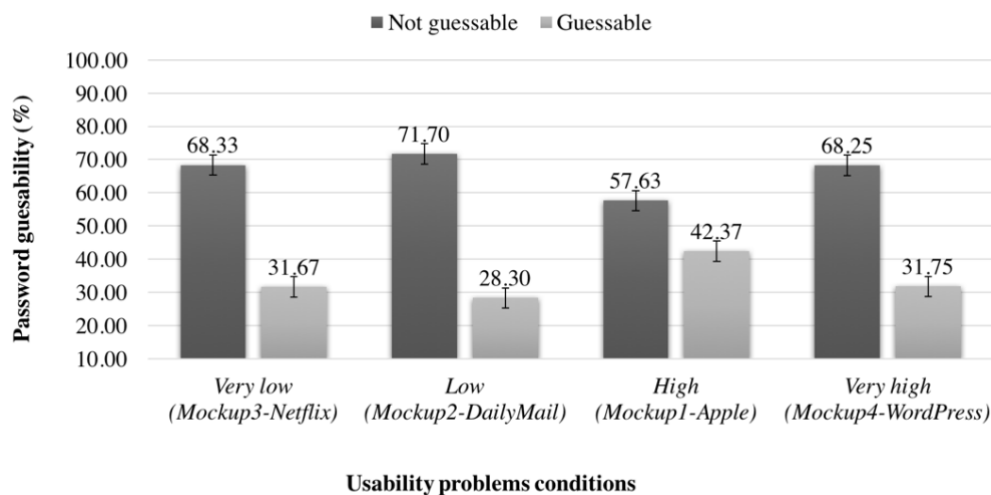


Figure 6.9 Percentages of password guessability across the four usability problems conditions

- ***Effect of policy presentation***

Figure 6.10 shows the percentage of password guessability for the four mockup PCSs conditions in terms of types of policy presentation. Looking at the policy presentation conditions individually, the results showed that there was a significant difference in all three conditions: before&after-interaction (*Mockup3-Netflix*: $x^2(1) = 8.07$, $p = .005$), during-interaction (*Mockup2-DailyMail*: $x^2(1) = 9.98$, $p = .002$), and during&after-interaction (*Mockup1-Apple* and *Mockup4-WordPress*: $x^2(1) = 8.39$, $p = .004$). The during-interaction (*Mockup2-DailyMail*) condition had the highest percentage of passwords that were not guessable (71.70%), whereas the during&after interaction (*Mockup1-Apple* and *Mockup4-WordPress*) had the highest percentage of guessable ones (36.90%).

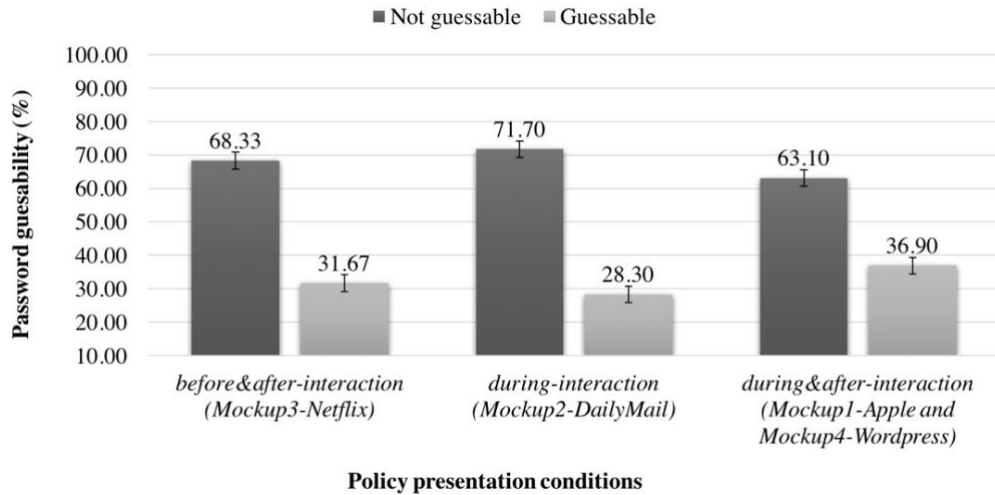


Figure 6.10 Percentages of password guessability across the three the policy presentation conditions. In summary, the strength of the created passwords did differ significantly between the four mockup PCS conditions in the number of digits, number of uppercase letters, and number of symbols included. Passwords included a significantly higher number of digits using *Mockup1-Apple*, a higher number of uppercase letters using *Mockup2-DailyMail* and *Mockup3-Netflix*, and a higher number of symbols using *Mockup4-WordPress*.

6.3.2 Password Recall

Out of 235, only 153 participants (65.11%) completed the recall task within 24 hours of the invitations being sent: 43 participants in the *Mockup1-Apple* condition, 31 in *Mockup2-DailyMail*, 38 in *Mockup3-Netflix*, and 41 in *Mockup4-WordPress*. Based on this sample, this section presents the results regarding the efficiency, effectiveness, and user satisfaction metrics between the four conditions.

The dependent measures for the recall part were not analysed in terms of the five factors discussed in the password creation (see Section 6.3.1), since the results remained the same among the individual factors.

6.3.2.1.1 Efficiency Metric

Recall time. Figure 6.11 shows the mean recall times for the four mockup PCS conditions. There was no significant difference in the recall time between the four conditions ($H(3) = 7.79, p = .051$).

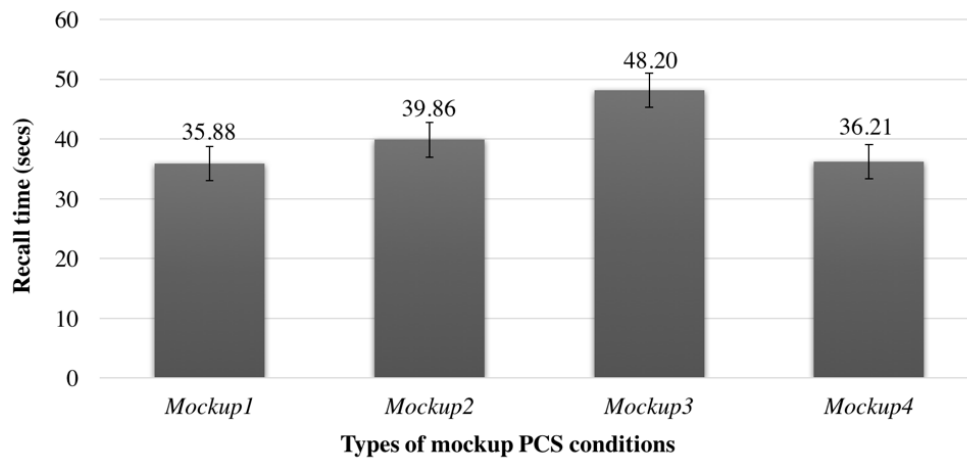


Figure 6.11 Mean recall times between the four mockup PCS conditions

6.3.2.1.2 Effectiveness Metric

Accuracy. Figure 6.12 shows the percentage of accuracy in recalling passwords for the four mockup PCS conditions. There was a significant difference in the accuracy of recalling passwords in the *Mockup2-DailyMail* condition ($\chi^2(1) = 3.90, p = .048$), but not in the other three conditions: *Mockup1-Apple* ($\chi^2(1) = 2.81, p = .093$), *Mockup3-Netflix* ($\chi^2(1) = 0.11, p = .746$), and *Mockup4-WordPress* ($\chi^2(1) = 0.61, p = .435$). In the *Mockup2-DailyMail* condition, 67.70% of passwords were successfully recalled, compared to 32.30% unsuccessful ones.

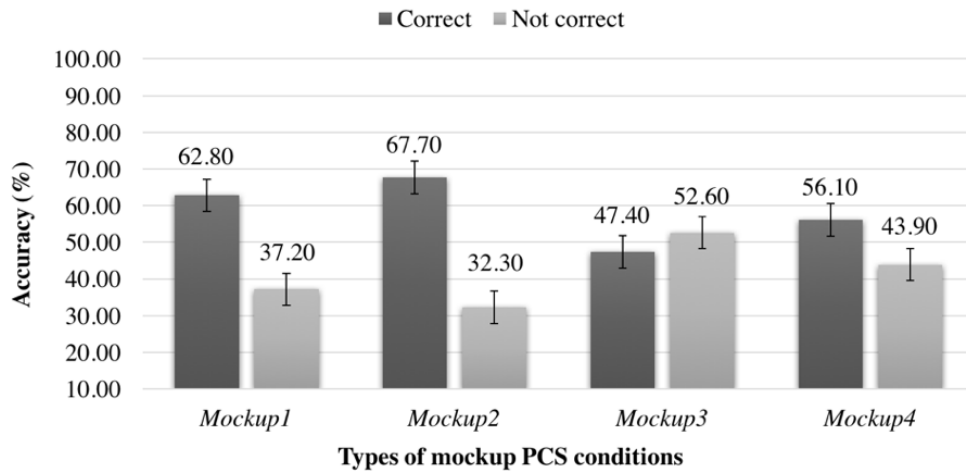


Figure 6.12 Percentages of accuracy in recalling passwords across the four mockup PCS conditions

6.3.2.1.3 User Satisfaction Metric

Confidence. Figure 6.13 shows the mean ratings of participants' confidence in recalling the correct passwords for the four mockup PCS conditions. There was no significant difference between mockup PCS conditions in participants' ratings of their confidence in recalling the correct passwords ($H(3) = 2.53, p = .470$).

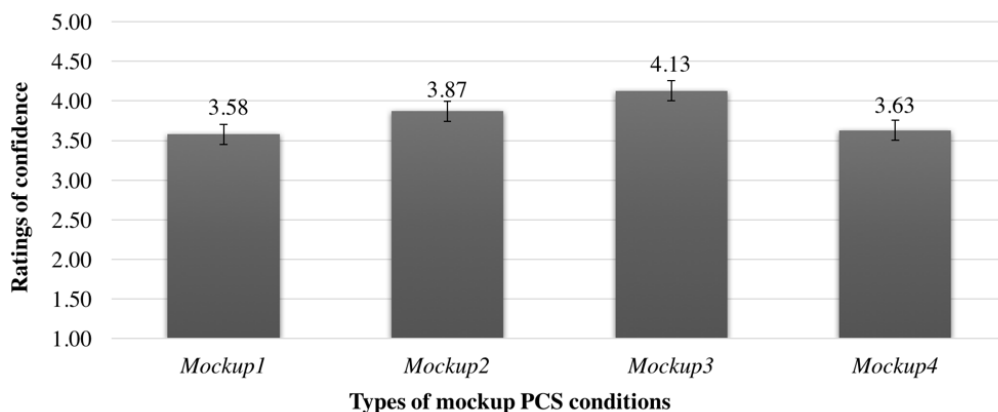


Figure 6.13 Mean ratings of participants' confidence between the four mockup PCS conditions

6.3.3 Users' Common Password Creation and Recall Practices

Participants reported that they had on average around 18.56 (standard deviation = 25.49) password-protected accounts and approximately 10.50 (standard deviation = 15.37) passwords. However, the majority of participants reported using the same password (167, 71.10%) or slightly different password (147, 62.60%) for multiple

accounts. When participants were asked about their actual behaviour in this study, very few reported creating their passwords based on reused ones (15, 6.38%) followed by modified ones (13, 5.53%). In contrast, the majority of participants reported creating entirely new passwords (207, 88.09%). It might be due to the fact that they were instructed to try their best in creating a new password for this study.

Regarding password change frequency, most participants (105, 44.68%) reported changing their passwords every three to six months, while very few of them (17, 7.2%) never changed them.

Many participants described themselves as being very knowledgeable (93, 39.60%) about what makes a secure password. Moreover, participants commented that secure passwords should have a combination of different character classes, and that they should not be based on personal information that might make them easy to guess. In contrast, others mentioned the length and the use of password managers as criteria for making secure passwords. Several participants (75, 31.90%) felt very confident about the strength of their most complicated password, and very few (3, 1.30%) did not feel confident at all.

Regarding password creation instructions, many participants (107, 45.50%) indicated that they ‘always’ read these instructions. However, there were circumstances when they did not do so, some of which were related to participants and others to the instructions themselves. The circumstances that related to participants were the following: (1) participants managed to create a password on their first attempt; (2) participants were familiar with instructions on the website as they were changing existing passwords and not creating new ones; and (3) participants were in a hurry and lacked time. The circumstances that related to the instructions themselves were: (1) the instructions were too lengthy; (2) the instructions were invisible in the PCS; and (3) the instructions were associated with low-value accounts.

Few participants (33, 14%) reported having a negative experience during the password creation process. They explained their frustration as follows: (1) PCSs enforced a very strict password policy (e.g. including uppercase letters and symbols in the password) that were not associated with high-value accounts, (2) it took a long time trying to

comply with the password policy, and (3) the chosen passwords were not allowed even after complying with the policy as they were similar to a password they had used in the past for the same account.

Furthermore, participants were asked about their familiarity with password management systems in which all passwords are stored, and only one password to that system must be remembered. Fewer participants had not heard of these systems (103, 43.80%) than those who had (132, 56.20%), but there was only a 12% difference between the two groups. Among those who had not heard of these systems, 57 (55.30%) were not interested in trying them. Among those who had heard of these systems, 78 (59.10%) were currently using one, 39 (29.50%) would never use one, and 15 (11.40%) had used one in the past but no longer did. They explained their reasons for ceasing their use as being memorability issues, password change frequency issues and, finally, trust concerns. Among those participants who returned to the recall task of the study, there were 36 (23.50%) who reported writing down the passwords they created. This behaviour seems persistent even with the existence of password management systems.

6.4 Discussion

This study aimed to investigate the effect of different PCSs' current practices on users and their passwords. To this end, four mockup PCS conditions (*Mockup1-Apple*, *Mockup2-DailyMail*, *Mockup3-Netflix*, and *Mockup4-WordPress*) were examined. The design of the mockup PCS conditions was based on the original design of four current PCSs that were previously evaluated with users (see Study 3, Chapter 5). The present study consisted of two parts: (1) password creation and (2) password recall; 235 users took part in the first part, while 153 returned for the second one. The first research question (RQ1) examined the differences between different mockup PCSs when users created passwords in terms of PCS usability and password strength. The second research question (RQ2) examined the same differences in terms of PCS usability, but when users recall passwords.

Regarding password creation, the efficiency of the PCSs and user satisfaction with them were both used to determine the usability level of the PCSs, whereas the password characteristic and guessability were used to examine the strength of the created passwords.

In order to have a better understanding of current practices, the four mockup PCSs were examined based on different factors: (1) number of usability problems, (2) whether and how they present the policy, (3) whether and how they present the suggestions, and (4) whether and how they present the strength indicators. Interestingly, the analysis of the different factors showed the same conclusion. For example, the passwords created with *Mockup2-DailyMail* and *Mockup3-Netflix* PCSs contained significantly more uppercase letters than the other mockup PCSs. In both ways of analyses, number of usability problems and policy presentation, this finding was supported. So, it was really hard to conclude whether this effect was caused by the usability level or policy presentation. Therefore, the effects of the current practices on the usability and password strength are discussed by the different mockup PCSs conditions.

In terms of the usability, the results revealed that the different mockup PCS conditions had an effect on the time to create passwords, but not on the keystrokes, perceived workload (except temporal demand), or user satisfaction when creating passwords.

Somewhat surprisingly, *Mockup4-WordPress* was found to be the most efficient design in terms of creation time (and keystrokes). Although *Mockup4-WordPress* had the highest number of usability problem, participants spent significantly less time creating passwords using this mockup. Thus, in contrast to expectations, the number of usability problems did not seem to affect the time to create passwords. For instance, *Mockup3-Netflix* had few usability problems, but the time spent creating passwords using it was longer than for *Mockup4-WordPress*. The reason for this is not clear, but it may relate to how the supporting features were structured and integrated in the PCSs.

However, there is also another possible explanation. *Mockup4-WordPress* was the only one that provided a show/hide password feature. This feature enabled users to visualise their passwords while entering them (i.e. unmasked passwords) instead of

showing the passwords in the form of asterisks (i.e. masked passwords). In fact, this feature was perceived as being unsafe during the user evaluation (see Study 3, Chapter 5), and was rated as a major problem. Thus, it seems that what users believed to be a usability problem might not in fact have affected the usability of the system. Even though in *Mockup4-WordPress* the password policy information was presented to the users for the first time during and after the password entry stage, there was a low number of keystrokes that resulted in short password creation time. One would expect that such a practice would require more time and make users hesitate as they were told about the policy while interacting with the PCS. However, it seems that the effect of this practice was minimised by providing the show/hide password feature. In other words, providing this feature might overcome the usability problems of PCSs.

In terms of password strength, the results showed that different mockup PCS conditions had an effect on password characteristics. In general, there was a significant difference in the password guessability ratio in each PCS condition. Passwords included a significantly higher number of digits using *Mockup1-Apple*, a higher number of uppercase letters using *Mockup2-DailyMail* and *Mockup3-Netflix*, and a higher number of symbols using *Mockup4-WordPress*. It is difficult to explain these results, but they might be related to how the policy or suggestion statements were phrased/constructed. For example, in the policy statements in both *Mockup2-DailyMail* and *Mockup3-Netflix*, the first element that was mentioned was the use of uppercase letters. Furthermore, in a similar vein, a set of possible symbols was provided in the suggestion statements in *Mockup4-WordPress*.

For the password recall, the results showed that different mockup PCS conditions had no effect on the recall time, the accuracy level (except *Mockup2-DailyMail*), and participants' confidence. Although there was no difference between the different PCSs, participants spent a longer time recalling passwords in *Mockup3-Netflix*, and the majority were not accurate.

These results must be interpreted with caution. As with much research on password behaviour, participants only imagined that they were creating a password for their online bank account, so although the use of a scenario was meant to increase validity,

the task did nevertheless lack ecological validity. However, the facts that 88.10% of participants reported creating an entirely new password and that 66.38% of the created passwords were not guessed by any of the five cracking approaches suggest that participants took the scenario and the password creation seriously. Furthermore, 23.50% of participants who returned for the second part of the study reported writing down their passwords, which suggests that they behaved in the same way they would normally do when managing their passwords.

6.5 Conclusions

To conclude, this study aimed to help us understand the current practices of PCSs and their effects on users and their passwords. The main finding suggests that current practices of PCSs had different effects on the usability of the mockup PCSs and the strength of passwords. More precisely, an effect was found on the time to create passwords and the characteristics of passwords. However, it is highly difficult to determine a specific practice that might have caused this effect. One reason is the interaction level between the supporting features and presentation timing was very high in the four mockup PCSs. Although a number of usability problems were deployed in the mockups, some had advantages in some aspect even with poor usability. Thus, the supporting features need to be examined individually and then combined with other supporting features to have a clear understanding of the impact of current PCS practices. This is discussed in Studies 6 and 7 in Chapters 8 and 9, respectively.

Phase 2: Testing Design Variables for PCSs

Chapter 7

Instructions for Creating Passwords: Analysis and User Study – *Study 5*

7.1 Introduction

User instructions in the form of password policies or password creation suggestions play a key role in helping users understand and comply with a PCS's requirements. However, current guidance provided by PCSs often seems ineffective for users when choosing passwords. In this vein, the results of Florencio and Herley's (2007) study indicated that users continue to tend to choose weak passwords.

Current practices of user instructions in PCSs are highly varied and often unclear. This may confuse users in a particular PCS, but also as they move from one PCS to another, which may in turn affect their password choice. The usability evaluations (see Table 5.9, Chapter 5), revealed that many of the usability problems identified regarding the password policies (60%) and creation suggestions (35%) were related to the amount of detail and clarity of the instructions. Therefore, it is important to examine user instructions in PCSs and ensure that they support users well when creating passwords.

User instructions can be classified into declarative information and procedural information (Ummelen, 1997). Declarative information is about facts, whereas procedural information is about actions. However, in the literature of user instructions, the use and effects of these two types of user instructions are not very clear (Karreman, Ummelen, & Steehouder, 2005).

For instance, Carroll and Mack (1984) concluded that user instructions have to be action-centred, since users tend to learn by doing and not by reading instructions. However, positive effects of declarative information have been found when forcing users to read this type of information (E. E. Smith & Goodman, 1984). Furthermore, Karreman et al.'s (2005) results indicated that reading declarative information positively affected task performance but negatively affected users' confidence. One way to solve this problem would be to provide both declarative and procedural information in user instructions. However, redundant information may result in higher cognitive load (Sweller & Chandler, 1991). Thus, it is important to provide the right type of information at the right time for users to perform their task successfully.

Therefore, the present study aims to provide a better understanding of the user instructions provided in PCSs by firstly examining the kinds of instructions that support users in creating passwords, and secondly investigating how the most frequently used instructions affect users' perceptions of the instructions. To address these aims, the study first analysed a total of 95 existing instructions to support users in creating passwords from a sample of 27 current PCSs. Based on this analysis, an online study was conducted with 117 respondents to understand how the most frequently used instructions affect users' perceptions of the instructions. The following four research questions are formulated to address these aims:

RQ1. What types of instructions are provided in current PCSs across the three timings of presentation?

RQ2. What is the most common used format (i.e. declarative or procedural) for each type of instruction in current PCSs across the three timings of presentation?

RQ3. Are there differences in the ratings of perceived helpfulness, clarity, amount of detail, and users' confidence between declarative and procedural format for each type of instruction across the three timings of presentation?

RQ4. Do the commonly used instructions match users' preferences for each type of instruction across the three timings of presentation?

7.2 Analysis of Current Instructions for Creating Passwords

A sample of current PCSs was analysed to understand what kinds of instructions are typically/commonly provided to support users in creating passwords. This aim addressed the first two research questions (RQ1 and RQ2).

7.2.1 Data Sources and Coding Scheme

A set of 27 websites with PCSs were selected from the top 100 entries on Alexa⁸ on 1 February 2016. Criteria for inclusion were that the website was in English, it had a dedicated PCS (i.e. it did not use other systems, such as Google or Facebook), and the PCS did not generate passwords automatically for users. Table 7.1 lists all the website PCSs analysed in this study, along with their descriptions and their Alexa ratings.

Table 7.1 PCSs analysed for this study

Website	Domain	Alexa rating
Adobe	Computer software company	91
Aliexpress	Online retailer	48
Amazon	Online retailer	6
Apple	Online retailer	52
BBC	Online newspaper	97
Disneystore	Online retailer	77
Dropbox	File hosting	95
eBay	Online auction site	23
Facebook	Social networking	2
GitHub	Development platform	82
Google	Search engine	1
Imgur	Online image hosting	43
IMDB	Online movie database	47

⁸ Alexa: <http://www.alexa.com/topsites>

Instagram	Social media	24
LinkedIn	Business-oriented social networking services	18
MSN	Web portal	14
MSN Office	Online office 365 application	71
Netflix	Online streaming media	30
Outbrain	Online advertiser	83
PayPal	Payment and money transfers service	39
Pinterest	Social networking	32
Reddit	Online social newspaper	36
Stackoverflow	Question and answer service for programmers	56
Tumblr	Micro-blogging platform and social networking	43
Twitter	Social networking and micro-blogging	10
WordPress	Blog web hosting	41
Yahoo	Search engine	5

These PCSs provided a range of different instructions at the three timings of presentation (i.e. before-interaction, during-interaction, and after-interaction), as discussed in Study 1 (see Section 3.3.2.1, Chapter 3), to guide users in creating new passwords. A total of 95 instructions were extracted, and their content was analysed. An open coding technique was used. Seven attributes emerged from the coding. The author and her supervisor coded all the instructions separately and together until there was complete agreement on the coding.

The seven attributes are the following.

- 1. Instruction type:** identified whether the instruction concerned password policy, a password creation suggestion, or an error message.
- 2. Explicit vs. implicit:** identified whether the instruction was given to the user as an explicit statement (*'Password needs at least one lowercase letter'*, GitHub) or an implicit statement (*'at least 6 characters'*, Amazon).
- 3. Declarative vs. procedural:** referred to the grammatical form of the instruction. Four grammatical forms were found: declarative, phrasal, modal, and imperative. The first three forms related to the declarative format, whereas

the last one related to the procedural format. Definitions and examples of the four forms are as follows:

- a. declarative statement is a declarative format which is expressed in the form of full sentence, e.g. (*'Good passwords are hard to guess'*, Dropbox);
- b. phrasal statement is a declarative format which is expressed in the form of group of words without a tensed verb, thus not a full sentence, e.g. (*'8 character minimum, case sensitive'*, Live);
- c. modal statement is a declarative format which is expressed in form of full sentence or phrase that includes *must* or *should*, e.g. (*'Must contain at least 1 more characters'*, Stackoverflow); and
- d. imperative statement is a procedural format which is expressed in the form of sentence which is a command, e.g. (*'Include at least 1 number or symbol (like !@#\$\$%^)'*, PayPal).

4. **Password-oriented vs. action-oriented:** whether the instruction was stated in language related to the password or to an action users should (not) take in creating their password. A password-oriented instruction is *'Short passwords are easy to guess'* (Google), whereas an action-oriented instruction is *'Avoid using the same password for multiple sites'* (eBay).
5. **General vs. specific:** the level of detail in the instruction. An example of a general instruction is *'Please create a password for your account'* (Disneystore). On the other hand, *'Your password is too short'* (Pinterest) is an example of a specific instruction.
6. **Positive vs. negative:** whether the instruction is positive or negative. Examples of negative instructions are: *'No consecutive identical characters'* (Outbrain) and *'Don't use a password from another site or something too obvious like your pet's name'* (Google).
7. **Polite command vs. brusque command:** the politeness element of the instruction. The instruction was considered to be polite if it contained the word 'please' (no other politeness forms were found in the instructions).

7.2.2 Results: Current State of Instructions for Creating Passwords

Figure 7.1 presents the temporal organization of the types of instruction across the three timings of presentation. Only 10% of instructions were provided at the before-interaction step (10, 10.52%). In contrast, nearly half of the instructions were presented at the during-interaction step of the PCS (45, 47.37%), and about 40% at the after-interaction step (40, 42.12%).

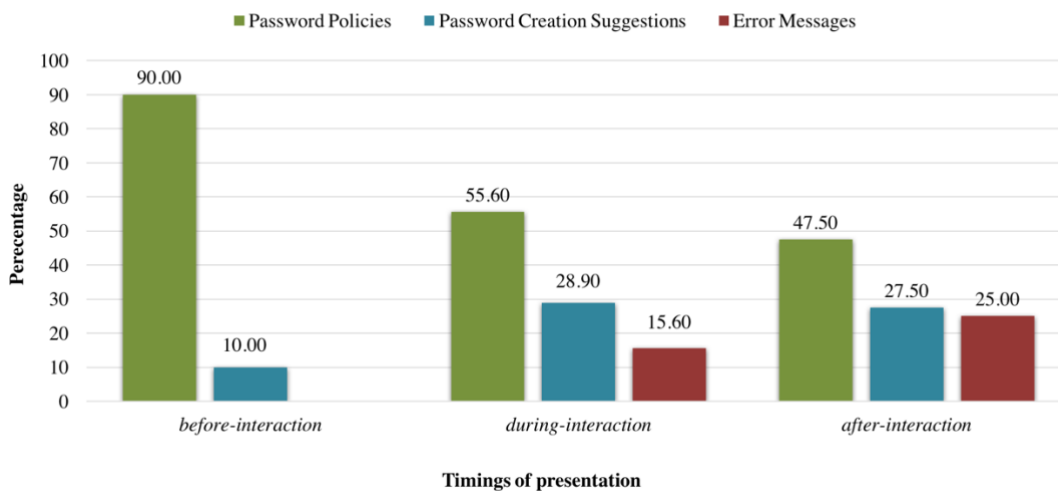


Figure 7.1 The percentage of the three presentation timings for the three types of instructions

7.2.2.1 Instructions at the Before-Interaction Step

As shown in Figure 7.1, a total of 10 password instruction statements were presented at the before-interaction step, including both password policies and password creation suggestions. However, 90.00% were password policies and only 10.00% were password creation suggestions. Thus, at this timing of presentation, PCSs typically support users only with information about what is needed and not about what makes a good password. Table 7.2 shows the frequency of instructions presented at the before-interaction step in relation to the seven attributes.

Regarding policy statements, most (7, 77.78%) were written implicitly. In addition, the use of the declarative format was twice as common (6, 66.67%) as the procedural format (3, 33.34%). Regarding the declarative format, both phrasal and declarative sentences were found. Finally, the one instance of a creation suggestion statement was written explicitly using a declarative format.

Table 7.2 Frequency of the instructions used at the before-interaction step across the seven attributes

<i>Instruction type</i>		Policy (<i>N=9</i>)	Creation suggestion (<i>N=1</i>)
<i>Explicit vs. Implicit</i>	Explicit	2	1
	Implicit	7	0
<i>Declarative vs. Procedural</i>	Declarative	<i>Declarative</i>	1
		<i>Phrasal</i>	5
		<i>Modal</i>	0
	Procedural	<i>Imperative</i>	3
<i>Password-oriented vs. Action-oriented</i>	Password-oriented	6	0
	Action-oriented	3	1
<i>General vs. Specific</i>	General	1	0
	Specific	8	1
<i>Positive vs. Negative</i>	Positive	9	1
	Negative	0	0
<i>Polite command vs. Brusque command</i>	Polite command	1	0
	Brusque command	8	1

In general, it was found that almost all instructions presented at the before-interaction step were specific and positive in format. Also, the politeness element did not occur at all in the procedural format.

7.2.2.2 Instructions at the During-Interaction Step

A total of 45 instruction statements were presented at the during-interaction step, including password policies, password creation suggestions, and error messages. Again, the password policies occurred the most frequently with 55.56%, followed by password creation suggestions with 28.89% and error messages with 15.56%, as shown in Figure 7.1. Table 7.3 presents the frequency of the instructions presented at the during-interaction step in relation to the seven attributes.

Of the policy statements, more than half (14, 56.00%) were written explicitly. The use of declarative policy statements was much more common (21, 84.00%) than procedural ones (4, 16.00%). The declarative policy statements mostly used modal sentences, followed by phrasal and declarative sentences. In addition, the policy statements were both positive and negative in both declarative and procedural formats

(except for the modal sentences, which were only stated positively). All instruction statements that related to policy were specific. Furthermore, the politeness element did not occur at all within the policy statements.

Table 7.3 Frequency of the instructions used at the during-interaction step across the seven attributes

<i>Instruction type</i>			Policy (<i>N</i> =25)	Creation suggestion (<i>N</i> =13)	Error message (<i>N</i> =7)
<i>Explicit vs. Implicit</i>	Explicit		14	10	5
	Implicit		11	3	2
<i>Declarative vs. Procedural</i>	Declarative	<i>Declarative</i>	6	5	7
		<i>Phrasal</i>	7	0	0
		<i>Modal</i>	8	0	0
	Procedural	<i>Imperative</i>	4	8	0
<i>Password-oriented vs. Action-oriented</i>	Password-oriented		21	5	7
	Action-oriented		4	8	0
<i>General vs. Specific</i>	General		0	6	4
	Specific		25	7	3
<i>Positive vs. Negative</i>	Positive		19	10	2
	Negative		6	3	5
<i>Polite command vs. Brusque command</i>	Polite command		0	3	0
	Brusque command		25	10	7

Regarding the creation suggestion statements, the majority (10, 76.92%) were written explicitly. Procedural suggestion statements were more common (8, 61.54%) than declarative ones (5, 38.46%). The declarative format of suggestions used only declarative sentences. When the suggestions were declarative, they were more often general than specific, but the frequency here is very small. In contrast, when suggestions were procedural, they were more often specific than general; however, again, the numbers are small. Negative constructions appeared only with the procedural suggestions. Finally, the politeness element only appeared with the positive procedural suggestions.

With regard to the error message statements, explicit statements were most common (5, 71.43%) than implicit ones (2, 28.57%). All error messages were written in the declarative format. The statements of the error messages tended to be more general than specific, but again, the frequency here is very small. The majority of the statements (5, 71.43%) were written using negative constructions. In addition, the politeness element did not occur at all within the error message statements.

7.2.2.3 Instructions at the After-Interaction Step

A total of 40 instruction statements presented at the after-interaction step were similar to the instructions types presented at the during-interaction step. Password policies occurred the most frequently with 47.50%, followed by password creation suggestions with 27.50% and error messages with 25.00%, as presented in Figure 7.1. Table 7.4 shows the frequency of the instructions presented at the after-interaction step in relation to the seven attributes.

Table 7.4 Frequency of the instructions used at the after-interaction step across the seven attributes

<i>Instruction type</i>		Policy (<i>N=19</i>)	Creation suggestion (<i>N=11</i>)	Error message (<i>N=10</i>)	
<i>Explicit vs. Implicit</i>	Explicit	14	11	7	
	Implicit	5	0	3	
<i>Declarative vs. Procedural</i>	Declarative	<i>Declarative</i>	1	0	8
		<i>Phrasal</i>	0	0	2
		<i>Modal</i>	12	0	0
	Procedural	<i>Imperative</i>	6	11	0
<i>Password-oriented vs. Action-oriented</i>	Password-oriented	12	0	10	
	Action-oriented	7	11	0	
<i>General vs. Specific</i>	General	0	9	7	
	Specific	19	2	3	
<i>Positive vs. Negative</i>	Positive	19	11	2	
	Negative	0	0	8	
<i>Polite command vs. Brusque command</i>	Polite command	4	8	1	
	Brusque command	15	3	9	

Of the policy statements, more than half (14, 73.68%) were written explicitly. Declarative policy statements (13, 68.42%) were more than twice as frequent as procedural ones (6, 31.58%). For the declarative policy statements, modal sentences were much more common (12, 92.30%) than declarative sentences (1, 7.69%). In addition, the policy statements were only written using positive wording. All instruction types that related to policy were specific. Finally, the politeness element was used frequently with the procedural policy statements, although frequency was small.

Regarding the creation suggestion statements, all were written explicitly in the procedural format. General suggestion statements were much more common than specific ones. All statements were presented positively. Also, nearly three-quarters of them used the politeness element.

Finally, for the error message statements, explicit messages were more common than implicit ones. Again, all error message statements were written in the declarative format, and the use of declarative sentences was more common than that of phrasal sentences. In addition, general error message statements (7, 70%) were more common than specific ones (3, 30%). The majority of the error message statements (8, 80%) were written negatively, and only one statement included the politeness element.

7.3 User Study on Instructions for Creating Passwords

An online user study was conducted to (1) understand how the most frequently used instructions affect users' perceptions of the instructions and (2) investigate what forms of instructions users prefer for the statements of password policy, password creation suggestions, and error messages across the three timings of presentation. These aims addressed the last two research questions stated at the beginning of the chapter (RQ3 and RQ4). The user study was conducted using an online questionnaire in which respondents were asked to rate and comment on a number of different possible instructions in the context of imagining creating a password.

7.3.1 Method

7.3.1.1 Design

This study had one independent variable, which was the format of the instruction, with two conditions: declarative and procedural. For each timing of presentation, the three types of instruction (i.e. password policy, password creation suggestion, and error message) were examined using variations of declarative and procedural statements. The design was not a full factorial one, as error message type, was not examined at the before-interaction step and was considered in the declarative format only. The reason for not including the error messages at the before-interaction step was that at this stage, users have not yet started interacting with the PCSs, so they cannot have experienced any failed attempts. Furthermore, error messages were only considered in the declarative format because these statements tend to be informative instead of procedural.

A total of 50 instructions statements were investigated in the user study. Due to this large number, the choice was made to divide the statements between three groups of respondents. Each group answered a questionnaire that had between 14 and 18 different instruction statements and took approximately 20-25 minutes to complete. The division into groups was not meant to create a between-group comparison but to accommodate the large number of instructions under investigation. The user instruction statements were investigated in the context of an imaginary online service that included a PCS. Respondents were asked to imagine creating a new password using this PCS with the help of the instructions provided. Table 7.5 illustrates the number of statements for each instruction type with the timing of presentation for each group.

Each respondent in Group 1 was shown 14 statements of password policy at the before-interaction and during-interaction steps. Each respondent in Group 2 was shown 18 statements of password creation suggestions at the before-interaction and during-interaction steps. Each respondent in Group 3 was shown 18 error message

statements at the during-interaction step; and password policy, password creation suggestion, and error message statements at the after-interaction step.

Table 7.5 Number of statements examined for each type of instruction across the three timings of presentation in each group

		Group 1	Group 2	Group 3
<i>before-interaction</i> <i>step</i>	Policy	5	-	-
	<i>Declarative</i>	2		
	<i>Procedural</i>	3		
	Creation suggestion	-	9	-
	<i>Declarative</i>		4	
	<i>Procedural</i>		5	
<i>during-interaction</i> <i>step</i>	Policy	9	-	-
	<i>Declarative</i>	6		
	<i>Procedural</i>	3		
	Creation suggestion	-	9	-
	<i>Declarative</i>		3	
	<i>Procedural</i>		6	
<i>after-interaction</i> <i>step</i>	Error message	-	-	3
	<i>Declarative</i>			3
	Policy	-	-	6
	<i>Declarative</i>			3
	<i>Procedural</i>			3
	Creation suggestion	-	-	4
	<i>Declarative</i>			1
	<i>Procedural</i>			3
	Error message	-	-	5
	<i>Declarative</i>			5
Total		14	18	18

The following measurements were used as the dependent variables: respondents' ratings of the (1) perceived helpfulness of instruction, (2) perceived clarity of instruction, (3) perceived level of detail of instruction, and (4) confidence in creating a password after reading the instruction. The four dependent variables were measured using 5-point Likert items ranging from 1 (not at all helpful/ not at all clear/ far too little detail/ not at all confident) to 5 (extremely helpful/ extremely clear/ far too much

detail/ extremely confident). In an optional question, participants were also offered the chance to explain their ratings for each instruction.

7.3.1.2 Respondents

A total of 228 respondents took part in the study. However, 111 were excluded for not completing the questionnaire or for doing so in less than 3 minutes (meaning they had not taken it seriously), or for giving identical answers for all statements. This left a total of 117 respondents to include in the analysis. The criteria of exclusion have been applied carefully to improve the data quality. For more details on the handling and justification of data exclusion, see Section 11.3 in Chapter 11.

The respondents were recruited from the University of York, Social Network sites, and MTurk crowdsourcing platform. The recruitment methods were varied to increase the range of respondents and the number of participants per group, and to balance the sample size across the three groups.

Group 1 included 40 respondents, all of whom were recruited from the Department of Computer Science at the University of York. Group 2 consisted of 15 respondents (non-MTurkers) from the Department of Theatre, Film and Television and the Department of Management and Law at the University of York, and 19 respondents (MTurkers) from MTurk. In Group 3, 18 respondents (non-MTurkers) were recruited from Social Network sites and 25 respondents (MTurkers) from MTurk.

For Groups 2 and 3, the difference between non-MTurkers and MTurkers was tested, and no significant differences were found (see Appendix C, Section C.1 for the statistical analyses). Therefore, the data from non-MTurkers and MTurkers in each group were combined for further analysis.

Respondents in Group 1 voluntarily participated in the study, while for Groups 2 and 3, MTurkers were compensated in the form of USD 0.50 (GBP 0.40), and non-MTurkers were entered in a prize draw of 10 Amazon vouchers worth GBP 10 each.

Table 7.6 Demographic characteristics (frequency and %) of respondents in each group and overall

Characteristics		Groups of participants			Overall (N=117)
		Group 1 (N=40)	Group 2 (N=34)	Group 3 (N=43)	
Gender	<i>Female</i>	16 (40)	13 (38.24)	20 (46.51)	49 (41.88)
	<i>Male</i>	24 (60)	21 (61.76)	23 (53.48)	68 (58.12)
Language	<i>English</i>	25 (62.50)	30 (88.24)	37 (86.05)	92 (78.63)
	<i>Other</i>	15 (37.50)	4 (11.76)	6 (13.95)	25 (21.37)
Education	<i>School</i>	1 (2.50)	5 (14.70)	1 (2.33)	7 (5.98)
	<i>Diploma</i>	1 (2.50)	2 (5.88)	7 (16.28)	10 (8.55)
	<i>Bachelor's</i>	4 (10)	21 (61.76)	17 (39.53)	42 (35.90)
	<i>Master's</i>	16 (40)	5 (14.71)	13 (30.23)	34 (29.06)
	<i>Doctoral</i>	18 (45)	1 (2.94)	5 (11.63)	24 (20.51)
Major/ Career	<i>Computing</i>	30 (75)	6 (17.65)	20 (46.51)	56 (47.86)
	<i>Non-computing</i>	10 (25)	28 (82.35)	23 (53.49)	61 (52.14)

Table 7.6 summarises the demographic characteristics of the respondents per group and overall. In total, 49 (41.88%) were females and 68 (58.12%) males. They ranged in age from 19 to 68 years, with a mean age of 36.33 years (standard deviation = 12.45). A majority of respondents (92, 78.63%) were native speakers of English, the remaining had been speaking English for on average 17.77 years (standard deviation = 10.89). Almost half of the respondents (58, 49.57%) had a postgraduate degree. The level of education of the remaining respondents ranged from bachelor's degree (42, 35.9%) to school degree (7, 5.98%). In general, the respondents' majors/career backgrounds were divided evenly between computing (56, 47.86%) and non-computing (61, 52.14%) fields. On average, the majority of respondents spent more than 6 hours a day online and using computers. The respondents' characteristics were almost the same between the three groups, except for the education level and major/career background. Most of the respondents in Groups 2 and 3 had a bachelor's degree. Moreover, most of the respondents in Group 2 were from non-computing fields.

7.3.1.3 Materials

This section presents the design of the questionnaire used for each group. It also covers the user instruction statements and their development. Table 7.7 illustrates the overall structure of the questionnaire.

Table 7.7 The overall structure of the questionnaire in each group

Section of the questionnaire	Group 1	Group 2	Group 3
Briefing about the study	This section was the same across the three groups		
<i>Timing of presentation</i> <i>Type of instruction</i> - Introduction page - Set of statements: respondents had to rate each statement using 5-point Likert items on four dependent measures	The section was repeated twice: <i>1. policy before interaction</i> <i>2. policy during interaction</i>	The section was repeated twice: <i>1. suggestion before interaction</i> <i>2. suggestion during interaction</i>	The section was repeated four times: <i>1. error message during interaction</i> <i>2. policy after interaction</i> <i>3. suggestion after interaction</i> <i>4. error message after interaction</i>
Post-study questions	This section was the same across the three groups		

7.3.1.3.1 Questionnaire Design

Three questionnaires were developed in this study, one for each group. However, there were common characteristics between the questionnaires. First, all questionnaires had the same structure: they began with a briefing that covered the overall purpose of the study, and they ended with post-study questions. As shown in Table 7.7, the timing of presentation and the type of instruction factors were used to split the set of statements within each group.

Imagine you are about to start creating a new password for an online service. **The only requirements are that your password have at least six characters and at least one numeral.**

You will see a number of different instructions that the service might provide before you start entering your password to help you create an appropriate password.

Figure 7.2 An example of the introduction page that presented the policy instruction at the before-interaction step in Group 1

(a) PCS presents a policy statement at the *before-interaction* step

(b) PCS presents a policy statement at the *during-interaction* step; the cursor was placed inside the new password entry field to demonstrate the password entry step

(c) PCS presents a policy statement at the *after-interaction* step; the cursor was placed inside the confirm password entry field to demonstrate the step after entering the password

Figure 7.3 Examples of the PCS images presented for a password policy statement across the three timings of presentation

Before presenting each set of statements, an introduction page was provided. This page described the timing of presentation that respondents were about to experience and the user instructions for creating a password using the PCS. Figure 7.2 shows an example of the introduction page that was presented at the before-interaction step for the policy instruction. The same information was provided for the remaining presentation timings except for the first two lines, which were replaced by the following: ‘*Imagine now you are actually entering a new password for an online service. The only requirement/suggestion is that.....*’ for the during-interaction step; and ‘*Imagine you entered a new password for an online service. The only requirement/suggestion is that.....*’ for the after-interaction step. In addition, an image of PCS with the user instruction statement was provided for each statement respondents received to help them visualise the PCS with the timing of presentation. Figure 7.3 shows examples of

the images presented for a password policy statement across the three timings of presentation.

7.3.1.3.2 Statements of Instruction and their Development

The 50 instruction statements were derived from the analysis of current PCS practices of using instructions (see Section 7.2.2). Figure 7.4 shows an example of how the policy statements for the before-interaction step were developed from the analysis.

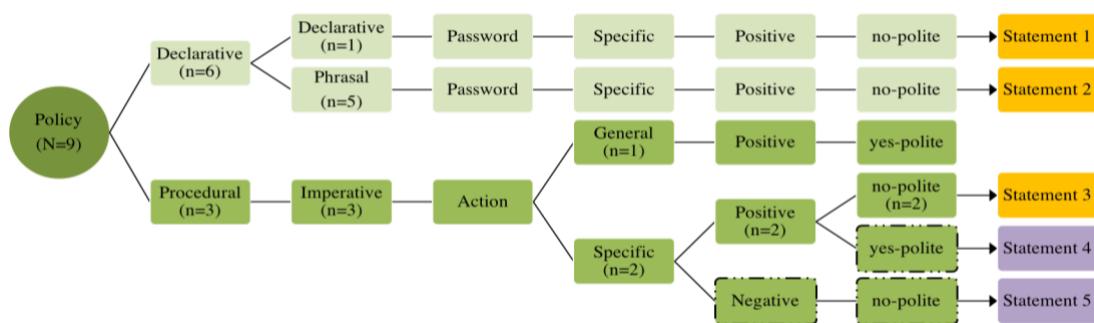


Figure 7.4 An example of the policy statements that were investigated in the user study and how they were derived from the analysis of the instructions currently used at the before-interaction step

As shown in Figure 7.4, both declarative and procedural statements are used in current PCSs to present policy instructions at the before-interaction step. Therefore, both were investigated in this user study to see which form was preferred. A negative format of policy statements is not used in current PCSs, but the choice was nevertheless made to include this type of format of procedural statements to investigate its effect. Finally, it was decided to include only specific policy statements in the user study since the general ones used in the current PCSs were not that helpful (e.g. ‘*please create a password for your account*’, DisneyStore).

Therefore, five instruction statements were developed for the before-interaction step: two declarative statements and three procedural statements. The two variations of the declarative statements were declarative-positive (Statement 1) and phrasal-positive (Statement 2), whereas the three variations of the procedural statement were imperative-positive (Statement 3), imperative-positive-polite (Statement 4), and imperative-negative (Statement 5). As shown in Figure 7.4, Statement 1, 2, and 3, did

appear in the analysis and were considered for the user study, whereas Statement 4 and 5 did not appear in the analysis, but were considered in the user study.

Table 7.8 Nature of the user instruction for each type across the three steps of interaction in each group, and the number of statement variations

	Timing of presentation	Type of instruction	Nature of user instruction	No. of statements
Group 1	<i>Before interaction</i>	Policy	The policy of the PCS is to have 6 characters and 1 numeral in the password	5
	<i>During interaction</i>	Policy	The policy of the PCS is to have only lowercase letters in the password	9
Group 2	<i>Before interaction</i>	Creation suggestion	The suggestion provided in the PCS is to not include uncommon words in the password (abstract), and to have both letters and numbers in the password (concrete).	9
	<i>During interaction</i>	Creation suggestion	The suggestion provided in the PCS is to add jokes in the password (abstract); and to add symbols in the password (concrete)	9
Group 3	<i>During interaction</i>	Error message	The error message provided in the PCS is that the password should be strong and uncommon.	3
	<i>After interaction</i>	Policy	The policy of the PCS is to have a combination of uppercase letters, lowercase letters, and symbols in the password.	6
		Creation suggestion	The suggestion provided in the PCS is to have at least eight characters (concrete).	4
		Error message	The error message provided in the PCS is that the password should be long and non-guessable.	5

The same procedure was carried out for all types of instructions for each timing of presentation to develop the instruction statements for this study. A very large number of instruction statement constructions could have been investigated in the study. Therefore, the researcher and her supervisor used their judgement to some extent in choosing statements for investigation.

Table 7.8 shows that nature of the user instruction for each type across the three steps of interaction in each group. As shown in Table 7.8, the timing of presentation and type of instruction were used to split the statements into sets within each group in terms of the nature of the user instructions.

For Group 1, the questionnaire consisted of 14 statements of password policy which are appropriate for the before-interaction and during-interaction steps. All policy statements at the before-interaction step used the following policy: ‘password must have at least six characters and at least one numeral’. Five variations were created to present this policy: two declarative statements (declarative-positive and phrasal-positive) and three procedural statements (imperative-positive, imperative-positive-polite, and imperative-negative). For the during-interaction step, the policy statements addressed the following policy: ‘password must have lowercase letters only’. There were nine variations of stating this policy: six declarative statements (declarative-positive, declarative-negative, modal-positive, modal-negative, phrasal-positive, and phrasal-negative) and three procedural statements (imperative-positive, imperative-positive-polite, and imperative-negative). Examples of the policy statements used at the before-interaction and during-interaction steps are:

1. The password needs to have at least six characters and at least one numeral (*declarative policy (declarative-positive) at the before-interaction step*);
2. Please use at least six characters and at least one numeral (*procedural policy (imperative-positive-polite) at the before-interaction step*);
3. The password must have only lowercase letters (*declarative policy (modal-positive) at the during-interaction step*);
4. Do not use uppercase letters (*procedural policy (imperative-negative) at the during-interaction step*).

For Group 2, the questionnaire consisted of 18 statements of password creation suggestions which are appropriate for the before-interaction and during-interaction steps. For the before-interaction step, all the statements had the following suggestion: ‘use both letters and numbers or only uncommon words’. There were nine variations of this suggestion: four declarative statements (declarative-password-abstract, declarative-password-concrete, declarative-action-abstract, declarative-action-concrete) and five procedural statements (imperative-positive-abstract-how&why, imperative-positive-concrete-how&why, imperative-positive-abstract-how, imperative-positive-abstract-how-polite, and imperative-negative-abstract-how).

Then, for the during-interaction step, the statements provided the following suggestion: ‘the password will be better if you add symbols or jokes’. Nine variations were created: three declarative statements (declarative-general, declarative-specific-abstract, and declarative-specific-concrete) and six procedural statements (imperative-general, imperative-general-polite, imperative-specific-abstract, imperative-specific-concrete, imperative-specific-abstract-polite, and imperative-specific-negative). The following are examples of the creation suggestion statements used at the two steps:

1. Good passwords have uncommon words (*declarative suggestion (declarative-password-abstract) at the before-interaction step*);
2. Use both letters and numbers to make a good password (*procedural suggestion (imperative-positive-concrete-how&why) at the before-interaction step*);
3. You can improve your password by adding jokes (*declarative suggestion (declarative-specific-abstract) at the during-interaction step*);
4. Add symbols to make your password stronger (*procedural suggestion (imperative-specific-concrete) at the during-interaction step*).

For Group 3, the questionnaire consisted of 18 error message statements which are appropriate for the during-interaction step, and password policy, creation suggestion, and error message statements at the after-interaction step. For the during-interaction step, all error message statements had the following meaning: ‘the password is strong and uncommon’. There were three declarative variations (declarative-general-positive, declarative-general-negative, declarative-specific-negative). For the after-interaction step, six policy statements, four suggestion statements, and five error message statements were used. For the policy statements, the following policy was used: ‘the password should have a combination of uppercase, lowercase, and symbols’. Three declarative statements (declarative-positive, modal-positive and modal-negative) and three procedural statements (imperative-positive, imperative-positive-polite, and imperative-negative) were created. The suggestion statements provided the following advice: ‘the password has at least eight characters’, with one declarative statement (declarative-specific-concrete) and three procedural statements (imperative-general, imperative-general-polite, and imperative-specific). Finally, the error message statements had the following meaning: ‘the password should be long

and non-guessable'. There were five declarative variations (declarative-general-negative, declarative-general-negative-polite, declarative-general-positive, declarative-specific-negative, and phrasal-general-negative). The following are examples of the statements of the three types of instructions used in the during- and after-interaction condition:

1. This is a very common password (*declarative error message (declarative-general-positive) at the during-interaction step*);
2. The password should be a combination of uppercase letters, lowercase letters, and symbols (*declarative policy (modal-positive) at the after-interaction step*);
3. Do not use only uppercase letters, lowercase letters, and symbols (*procedural policy (imperative-negative) at the after-interaction step*);
4. Good passwords have at least eight characters (*declarative suggestion (declarative-specific-concrete) at the after-interaction step*);
5. Choose a more secure password (*procedural suggestion (imperative-general) at the after-interaction step*);
6. Your password is too short (*declarative error message (declarative-specific-negative) at the after-interaction step*).

The order in which the declarative and procedural statements were presented was counterbalanced between respondents for each type of instruction across the timing of presentation. A full list of the 50 instruction statements, indicating the timing of presentation and the type of instruction for each group, is provided in Appendix C, Section C.2.

7.3.1.3.3 Post-Study Questions

At the end of the questionnaire, a post-study questions page was provided. The questions were split into two parts and were similar to the ones used in Study 4 (See Section 6.2.3.1.3, Chapter 6), as follows:

- Information about respondents' password-related behaviours: this part asked about the respondents' approximate number of online protected accounts and total number of passwords. It also included questions about respondents'

password reuse and password-changing behaviours. In addition, respondents were asked about how frequently they read instructions when creating passwords, and about their knowledge regarding creating a secure password.

- Information about respondents' demographics: this part contained questions about respondents' age, gender, native language, education, major, and computer and internet usage.

7.3.1.4 Pilot of the Study Procedure

A pilot study was conducted with three postgraduate students from the Departments of Computer Science and Linguistics at the University of York to test the overall process and design of the study. The study procedure was perceived as being easy, and the instructions and tasks were clear to follow.

However, two participants raised a concern about missing the starting point of a new timing of presentation (a new set of statements). Therefore, for the main study, the researcher paid attention to this problem and added an introduction page (see Figure 7.2) to split up the set of statements and indicate the start of a new timing of presentation. Furthermore, another issue was raised regarding the PCSs that presented user instructions at the during-interaction and after-interaction steps. Originally, these PCSs were designed in GIF format: the user instructions were demonstrated in a live demonstration of a PCS to show the movement of the cursor between the password entry and confirm password entry fields, and when the user instructions appeared. However, this attempt was not successful because it caused confusion and did not illustrate the timing of presentation accurately. Therefore, for the main study, this issue was addressed by replacing the GIF format by static images (see Figure 7.3 b and c).

A second pilot study was then conducted with three other postgraduate students from the Department of Computer Science to make sure the changes addressed the concerns raised in the first pilot. No issues were reported this time, and it appeared that the improvements solved the issues. The data from the two pilot studies were not included in the results.

7.3.1.5 Procedure

The snowballing sampling method was used to distribute links to the three questionnaires via e-mailing lists and social network sites. The same links were also posted on the MTurk platform. A briefing about the study and an informed consent form were given at the beginning of the questionnaire (see Appendix A, Section A.1). Respondents were assured that they would not be asked to reveal any of their passwords or create any passwords during the study. Respondents were then asked to confirm their agreement and their understanding of the information provided in the briefing by clicking on the 'Next' button. After that, respondents were asked to imagine that they needed to create a new password using a PCS. The instruction statements were presented in the context of an imaginary online service that provided a PCS. Respondents were instructed to read the instructions provided in the PCS. Next, they were asked to rate the helpfulness, clarity, and amount of detail of the instruction; and their confidence level after reading it. Upon completion of the questionnaire, respondents were asked to answer the post-study questions.

7.3.1.6 Data Analysis

Kolmogorov-Smirnov and Shapiro-Wilk tests were used to test for normality on all dependent measures. All were significantly non-normal ($p < 0.05$), for both of these tests. Therefore, non-parametric statistics were used throughout the analysis. A set of within-participant analyses were conducted for each group to compare participants' performance between the two formats of instruction statements: declarative and procedural. Wilcoxon signed-ranks tests (Z statistic) were used for the within-participant analyses. Further analyses within the declarative and procedural statements were then conducted using Friedman's test (χ^2 statistic) if there were more than two variations of statements (i.e. K-related sample), and Wilcoxon signed-ranks tests (Z statistic) if there were two variations (i.e. 2-related sample).

7.3.2 Results

This section examines the respondents' answers regarding the two formats for the different types of instruction across the three timings of presentation. First, it presents

the policy and creation suggestion instructions provided at the before-interaction step, followed by the policy, creation suggestion, and error message instructions presented at the during-interaction step. Finally, it presents the three types of instruction provided at the after-interaction step.

7.3.2.1 Instructions at the Before-Interaction Step

7.3.2.1.1 Policy Instructions

Table 7.9 shows the mean (and median) ratings for the perceived helpfulness, clarity, amount of detail, and respondents' confidence regarding the policy statements provided at the before-interaction step. Table 7.10 presents the results of the pairwise comparisons for the procedural policy statements across the dependent measures.

Overall, there were significant differences in the ratings of helpfulness ($Z = -2.56$, $p = .010$), clarity ($Z = -2.74$, $p = .006$), and confidence ($Z = -2.16$, $p = .031$) between the two formats of the policy instructions, but not in the ratings of amount of detail ($Z = -0.22$, $p = .827$). Respondents rated the helpfulness, clarity, and their confidence when reading declarative policy statements significantly higher than when reading procedural statements.

Table 7.9 Mean (median) ratings of the dependent measures: policy at the before-interaction step

	Helpfulness	Clarity	Amount of detail	Confidence
<i>Declarative policy statements vs. procedural policy statements</i>				
Overall declarative	3.78 (4.00)	3.75 (4.00)	2.70 (2.75)	3.93 (4.00)
Overall procedural	3.58 (3.67)	3.43 (3.50)	2.69 (2.83)	3.76 (3.67)
<i>p value</i>	.010	.006	n.s.	.031
<i>Declarative policy statements (2 variations)</i>				
declarative-positive	3.93 (4.00)	3.90 (4.00)	2.75 (3.00)	4.08 (4.00)
phrasal-positive	3.63 (4.00)	3.60 (4.00)	2.65 (3.00)	3.78 (4.00)
<i>p value</i>	.042	n.s.	n.s.	.037
<i>Procedural policy statements (3 variations)</i>				
imperative-positive	3.90 (4.00)	3.83 (4.00)	2.80 (3.00)	4.05 (4.00)
imperative-positive-polite	4.00 (4.00)	3.95 (4.00)	2.73 (3.00)	4.00 (4.00)
imperative-negative	2.83 (3.00)	2.53 (2.00)	2.55 (3.00)	3.23 (3.00)
<i>p value</i>	.000	.000	n.s.	.000

Between the two variations of declarative policy statements, there was a significant difference in the ratings of helpfulness ($Z = -2.03$, $p = .042$) and confidence ($Z = -2.09$,

$p = .037$), but not in the ratings of clarity ($Z = -1.74$, $p = .083$) and amount of detail ($Z = -1.00$, $p = .317$). For both helpfulness and confidence, respondents gave significantly higher ratings for the declarative-positive statement than the phrasal-positive statement.

Table 7.10 Pairwise comparisons between the procedural policy statements at the before-interaction step across the dependent measures

		imperative- positive	imperative- positive-polite	imperative- negative
Helpfulness	imperative-positive	-	-0.50	0.99*
	imperative-positive-polite		-	1.04*
	imperative-negative			-
Clarity	imperative-positive	-	-0.10	1.04*
	imperative-positive-polite		-	1.14*
	imperative-negative			-
Confidence	imperative-positive	-	0.06	0.80*
	imperative-positive-polite		-	0.74*
	imperative-negative			-

Note. * denotes a significant result in pairwise comparison, $p < .05$.

Between the three variations of procedural policy statements, there was a significant difference in the ratings of helpfulness ($\chi^2(2) = 43.39$, $p < .001$), clarity ($\chi^2(2) = 46.16$, $p < .001$), and confidence ($\chi^2(2) = 31.28$, $p < .001$), but not in the ratings of the amount of detail ($\chi^2(2) = 5.35$, $p = .069$). For both helpfulness and clarity, respondents gave significantly higher ratings for the imperative-positive-polite statement than for the other statements. While respondents felt more confident reading the imperative-positive statement than the other statements, a pairwise comparison showed no significant difference in the ratings of helpfulness, clarity, and confidence between the imperative-positive-polite and imperative-positive statements (see Table 7.10). On the other hand, there was a significant difference between the negative and positive formats of policy regardless of the politeness element.

7.3.2.1.2 Suggestion Instructions

Table 7.11 shows the mean (and median) ratings for the perceived helpfulness, clarity, amount of detail, and respondents' confidence regarding the suggestion statements provided at the before-interaction step; Table 7.12 and Table 7.13 present the results of the pairwise comparison for the declarative and procedural suggestion statements across the dependent measures, respectively.

Table 7.11 Mean (median) ratings of the dependent measures: suggestion at the before-interaction step

	Helpfulness	Clarity	Amount of detail	Confidence
<i>Declarative suggestion statements vs. procedural suggestion statements</i>				
Overall declarative	2.78 (2.75)	3.15 (3.25)	2.41 (2.50)	2.99 (3.00)
Overall procedural	2.39 (2.40)	2.75 (2.80)	2.08 (2.20)	2.54 (2.40)
<i>p value</i>	.000	.001	.000	.000
<i>Declarative suggestion statements (4 variations)</i>				
declarative-password-abstract	2.32 (2.00)	2.71 (2.50)	2.06 (2.00)	2.53 (2.00)
declarative-password-concrete	3.26 (3.50)	3.71 (4.00)	2.71 (3.00)	3.56 (4.00)
declarative-action-abstract	2.38 (2.00)	2.68 (2.50)	2.18 (2.00)	2.44 (2.00)
declarative-action-concrete	3.15 (3.00)	3.53 (4.00)	2.71 (3.00)	3.44 (4.00)
<i>p value</i>	.000	.000	.000	.000
<i>Procedural suggestion statements (5 variations)</i>				
imperative-positive-abstract-how&why	2.38 (2.00)	2.74 (3.00)	2.18 (2.00)	2.47 (2.50)
imperative-positive-concrete-how&why	3.38 (3.50)	3.85 (4.00)	2.71 (3.00)	3.71 (4.00)
imperative-positive-abstract-how	2.00 (2.00)	2.44 (2.00)	1.82 (2.00)	2.09 (2.00)
imperative-positive-abstract-how-polite	1.97 (2.00)	2.26 (2.00)	1.79 (2.00)	2.18 (2.00)
imperative-negative-abstract-how	2.21 (2.00)	2.44 (2.00)	1.91 (2.00)	2.24 (2.00)
<i>p value</i>	.000	.000	.000	.000

Overall, there were significant differences in ratings of helpfulness ($Z = -3.63$, $p < .001$), clarity ($Z = -3.30$, $p < .001$), amount of detail ($Z = -4.12$, $p < .001$), and confidence ($Z = -4.07$, $p < .001$) between the two formats of the suggestion instructions. Respondents rated the helpfulness, clarity, amount of detail, and their confidence significantly higher for the declarative suggestion than for the procedural one.

For the declarative suggestion statements, there were significant differences in ratings on the four dependent measures between the four variations: helpfulness ($\chi^2(3) = 33.05$, $p < .001$), clarity ($\chi^2(3) = 35.60$, $p < .001$), amount of detail ($\chi^2(3) = 39.63$, $p < .001$), and confidence ($\chi^2(3) = 33.05$, $p < .001$).

.001), and confidence ($\chi^2(3) = 43.79, p < .001$). Respondents gave significantly higher ratings on the four dependent measures for the declarative-password-concrete statement than for the other statements. However, the pairwise comparison showed a significant difference on the four dependent measures between the abstract and concrete types of declarative suggestion regardless of whether the suggestion was written using the password- or action-oriented format (see Table 7.12). Thus, the concrete suggestions were always rated significantly higher than the abstract ones.

Table 7.12 Pairwise comparisons between the declarative suggestion statements at the before-interaction step across the dependent measures

		declarative- password- abstract	declarative- password- concrete	declarative- action- abstract	declarative- action- concrete
Helpfulness	declarative-password-abstract	-	-1.12*	-0.10	-1.13*
	declarative-password-concrete		-	1.02*	-0.02
	declarative-action-abstract			-	-1.03*
	declarative-action-concrete				-
Clarity	declarative-password-abstract	-	-1.09*	0.04	-1.02*
	declarative-password-concrete		-	1.13*	0.07
	declarative-action-abstract			-	-1.06*
	declarative-action-concrete				-
Amount of detail	declarative-password-abstract	-	-1.13*	-0.18	-1.16*
	declarative-password-concrete		-	0.96*	-0.03
	declarative-action-abstract			-	-0.99*
	declarative-action-concrete				-
Confidence	declarative-password-abstract	-	-1.15*	0.10	-1.19*
	declarative-password-concrete		-	1.29*	0.44
	declarative-action-abstract			-	-1.25*
	declarative-action-concrete				-

Note. * denotes a significant result in pairwise comparison, $p < .05$.

Table 7.13 Pairwise comparisons between the procedural suggestion statements at the before-interaction step across the dependent measures

		imperative -positive- abstract- how&why	imperative -positive- concrete- how&why	imperative- positive- abstract- how	imperative- positive- abstract- how-polite	imperative -negative- abstract- how
Helpfulness	imperative-positive-abstract-how&why	-	-1.40*	0.66	0.77*	0.27
	imperative-positive-concrete-how&why		-	2.06*	2.16*	1.66*
	imperative-positive-abstract-how			-	0.10	-0.40
	imperative-positive-abstract-how-polite				-	-0.50
	imperative-negative-abstract-how					-
Clarity	imperative-positive-abstract-how&why	-	-1.28*	0.46	0.79*	0.47
	imperative-positive-concrete-how&why		-	1.74*	2.07*	1.75*
	imperative-positive-abstract-how			-	0.34	0.02
	imperative-positive-abstract-how polite				-	-0.32
	imperative-negative-abstract-how					-
Amount of detail	imperative-positive-abstract-how&why	-	-1.03*	0.62	0.75*	0.47
	imperative-positive-concrete-how&why		-	1.65*	1.78*	1.50*
	imperative-positive-abstract-how			-	0.13	-0.15
	imperative-positive-abstract-how polite				-	-0.28
	imperative-negative-abstract-how					-
Confidence	imperative-positive-abstract-how&why	-	1.59*	0.72	0.57	0.37
	imperative-positive-concrete-how&why		-	2.31*	2.16*	1.97*
	imperative-positive-abstract-how			-	-0.15	-0.35
	imperative-positive-abstract-how polite				-	-0.21
	imperative-negative-abstract-how					-

Note. * denotes a significant result in pairwise comparison, $p < .05$.

For the procedural suggestion statements, there were significant differences in ratings of helpfulness ($\chi^2(4) = 62.91$, $p < .001$), clarity ($\chi^2(4) = 53.09$, $p < .001$), amount of detail ($\chi^2(4) = 52.21$, $p < .001$), and confidence ($\chi^2(4) = 77.54$, $p < .001$) between the

five variations. Respondents gave significantly higher ratings for the imperative-positive-concrete-how&why statement on the four dependent measures compared to the other statements. Similar to the declarative suggestion statements, the pairwise comparison showed a significant difference between the abstract and concrete procedural suggestions on the four dependent measures (see Table 7.13). Furthermore, providing a negative or polite format of procedural suggestion appeared to have no effect on the four dependent measures as long as the nature of the suggestion was abstract.

The following summarises the results of the types of instruction presented at the before-interaction step. For both policy and suggestion statements, the perceived helpfulness, clarity, amount of detail, and users' confidence were higher for the declarative format than for the procedural one. For declarative policy statements, users preferred a positive declarative format, whereas for declarative suggestions, users preferred a concrete format.

7.3.2.2 Instructions at the During-Interaction Step

7.3.2.2.1 Policy Instructions

Table 7.14 shows the mean (and median) ratings for the perceived helpfulness, clarity, amount of detail, and respondents' confidence regarding the policy statements provided at the during-interaction step; Table 7.15 and Table 7.16 present the results of the pairwise comparison for the declarative and procedural policy statements for the procedural policy statements across the dependent measures, respectively.

Overall, there were significant differences in ratings of helpfulness ($Z = -4.41$, $p < .000$), clarity ($Z = -4.57$, $p < .001$), amount of detail ($Z = -3.93$, $p < .001$), and confidence ($Z = -4.12$, $p < .001$) between the two types of format of the policy instructions. In contrast to the results for type of policy presented at the before-interaction step, respondents gave significantly higher ratings on all four dependent measures for procedural policy than declarative policy statements.

Table 7.14 Mean (median) ratings of the dependent measures: policy at the during-interaction step

	Helpfulness	Clarity	Amount of detail	Confidence
<i>Declarative policy statements vs. procedural policy statements</i>				
Overall declarative	2.93 (3.00)	3.07 (3.17)	2.40 (2.50)	3.22 (3.17)
Overall procedural	3.38 (3.33)	3.63 (3.67)	2.63 (2.67)	3.54 (3.67)
<i>p value</i>	.000	.000	.000	.000
<i>Declarative policy statements (6 variations)</i>				
declarative-positive	3.25 (3.00)	3.40 (3.50)	2.53 (3.00)	3.45 (3.50)
declarative-negative	2.28 (2.00)	2.10 (2.00)	2.18 (2.00)	2.80 (3.00)
modal-positive	3.45 (3.00)	3.60 (4.00)	2.58 (3.00)	3.68 (4.00)
modal-negative	2.78 (3.00)	2.93 (3.00)	2.28 (2.00)	2.95 (3.00)
phrasal-positive	3.13 (3.00)	3.38 (3.00)	2.55 (3.00)	3.45 (3.00)
phrasal-negative	2.73 (3.00)	3.03 (3.00)	2.28 (2.00)	2.98 (3.00)
<i>p value</i>	.000	.000	.049	.000
<i>Procedural policy statements (3 variations)</i>				
imperative-positive	3.53 (4.00)	3.88 (4.00)	2.78 (3.00)	3.65 (4.00)
imperative-positive-polite	3.68 (4.00)	3.95 (4.00)	2.75 (3.00)	3.85 (4.00)
imperative-negative	2.93 (3.00)	3.08 (3.00)	2.38 (3.00)	3.13 (3.00)
<i>p value</i>	.000	.000	.001	.000

For the declarative policy statements, there were significant differences in ratings on the four dependent measures between the six variations: helpfulness ($\chi^2(5) = 51.87$, $p < .001$), clarity ($\chi^2(5) = 52.99$, $p < .001$), amount of detail ($\chi^2(5) = 11.14$, $p = .049$), and confidence ($\chi^2(5) = 30.58$, $p < .001$). Respondents gave significantly higher ratings on the four measures for the modal-positive statement than for the other statements. Furthermore, the pairwise comparison showed a significant difference between the negative and positive formats for the declarative policy and modal policy statements (see Table 7.15). The positive format was always rated higher than the negative one.

Table 7.15 Pairwise comparisons between the declarative policy statements at the during-interaction step across the dependent measures

		declarative- positive	declarative- negative	modal- positive	modal- negative	phrasal- positive	phrasal- negative
Helpfulness	declarative-positive	-	1.85*	-0.28	0.90*	0.29	0.91*
	declarative-negative		-	-2.13*	-0.95*	-1.56*	-0.94*
	modal-positive			-	1.18*	0.56	1.19*
	modal-negative				-	-0.61	0.01
	phrasal-positive					-	0.63
	phrasal-negative						-
Clarity	declarative-positive	-	2.10*	-0.11	0.78	0.29	0.70
	declarative-negative		-	2.21*	-1.33*	-1.81*	-1.40*
	modal-positive			-	0.89*	0.40	0.81
	modal-negative				-	-0.49	-0.08
	phrasal-positive					-	0.41
	phrasal-negative						-
Amount of detail	declarative-positive	-	0.70	-0.09	0.45	-0.01	0.45
	declarative-negative		-	-0.79	-0.25	-0.71	-0.25
	modal-positive			-	0.54	0.08	-0.54
	modal-negative				-	-0.46	0.00
	phrasal-positive					-	0.46
	phrasal-negative						-
Confidence	declarative-positive	-	1.11*	-0.40	-0.46	-0.03	0.63
	declarative-negative		-	-1.58*	-0.49	-1.14*	-0.49
	modal-positive			-	1.09*	0.44	1.09*
	modal-negative				-	-0.65	0.00
	phrasal-positive					-	0.65
	phrasal-negative						-

Note. * denotes a significant result in pairwise comparison, $p < .05$.

For the procedural policy statements, there were significant difference in ratings on the four dependent measures between the three variations: helpfulness ($\chi^2(2) = 16.14$, $p < .001$), clarity ($\chi^2(2) = 18.18$, $p < .001$), amount of detail ($\chi^2(2) = 14.98$, $p = .001$), and confidence ($\chi^2(2) = 15.28$, $p < .001$). Similar to the procedural policy statements provided at the before-interaction step, respondents gave significantly higher ratings on the helpfulness, clarity, and confidence for the imperative-positive-polite statement than for the other statements, whereas they gave higher ratings on the amount of detail for the imperative-positive statement. However, the pairwise comparison showed no significant difference in the ratings of the four dependent measures between the imperative-positive-polite and imperative-positive statements (see Table 7.16). On the

other hand, there was a significant difference between the negative and positive formats of procedural policy statements.

Table 7.16 Pairwise comparisons between the procedural policy statements at the during-interaction step across the dependent measures

		imperative- positive	imperative- positive-polite	imperative- negative
Helpfulness	imperative-positive	-	-0.16	0.43
	imperative-positive-polite		-	0.59*
	imperative-negative			-
Clarity	imperative-positive	-	-0.06	0.55*
	imperative-positive-polite		-	0.61*
	imperative-negative			-
Amount of detail	imperative-positive	-	0.05	0.44*
	imperative-positive-polite		-	0.39
	imperative-negative			-
Confidence	imperative-positive	-	-0.21	0.40
	imperative-positive-polite		-	0.61*
	imperative-negative			-

Note. * denotes a significant result in pairwise comparison, $p < .05$.

7.3.2.2.2 Suggestion Instructions

Table 7.17 shows the mean (and median) ratings for the perceived helpfulness, clarity, amount of detail, and respondents' confidence regarding the suggestion statements provided at the during-interaction step; Table 7.18 and Table 7.19 present the results of the pairwise comparison for the declarative and procedural suggestion statements across the dependent measures, respectively.

Overall, there were significant differences in ratings of helpfulness ($Z = -4.23$, $p < .001$), clarity ($Z = -4.06$, $p < .001$), amount of detail ($Z = -2.96$, $p = .003$), and confidence ($Z = -4.35$, $p < .001$) between the two types of format of the suggestion instructions. Similar to the suggestions presented at the before-interaction step, respondents gave significantly higher ratings on all four dependent measures for declarative suggestions than for procedural suggestions.

Table 7.17 Mean (median) ratings of the dependent measures: suggestion at the during-interaction step

	Helpfulness	Clarity	Amount of detail	Confidence
<i>Declarative suggestion statements vs. procedural suggestion statements</i>				
Overall declarative	2.82 (2.67)	2.96 (3.00)	2.26 (2.33)	2.86 (2.67)
Overall procedural	2.35 (2.25)	2.48 (2.58)	1.98 (2.00)	2.33 (2.17)
<i>p value</i>	.000	.000	.003	.000
<i>Declarative suggestion statements (3 variations)</i>				
declarative-general	2.94 (3.00)	3.15 (3.00)	2.41 (3.00)	3.03 (3.00)
declarative-specific-abstract	1.74 (1.00)	1.97 (1.50)	1.68 (1.00)	1.79 (1.00)
declarative-specific-concrete	3.79 (4.00)	3.76 (4.00)	2.71 (3.00)	3.76 (4.00)
<i>p value</i>	.000	.000	.000	.000
<i>Procedural suggestion statements (6 variations)</i>				
imperative-general	2.06 (2.00)	2.24 (2.00)	1.85 (2.00)	2.06 (2.00)
imperative-general-polite	2.18 (2.00)	2.32 (2.00)	1.85 (2.00)	2.24 (2.00)
imperative-specific-abstract	1.88 (1.50)	1.97 (1.50)	1.68 (1.50)	1.79 (2.00)
imperative-specific-concrete	3.88 (4.00)	3.88 (4.00)	2.79 (3.00)	3.79 (4.00)
imperative-specific-abstract-polite	1.71 (1.00)	1.82 (1.50)	1.65 (1.00)	1.71 (1.00)
imperative-specific-negative	2.41 (2.50)	2.65 (3.00)	2.07 (2.00)	2.38 (3.00)
<i>p value</i>	.000	.000	.000	.000

Table 7.18 Pairwise comparisons between the declarative suggestion statements at the during-interaction step across the dependent measures

		declarative-general	declarative-specific-abstract	declarative-specific-concrete
Helpfulness	declarative-general	-	0.85*	-0.63*
	declarative-specific-abstract		-	-1.49*
	declarative-specific-concrete			-
Clarity	declarative-general	-	0.82*	-0.47
	declarative-specific-abstract		-	-1.29*
	declarative-specific-concrete			-
Amount of detail	declarative-general	-	0.77*	-0.28
	declarative-specific-abstract		-	-1.04*
	declarative-specific-concrete			-
Confidence	declarative-general	-	0.97*	-0.49*
	declarative-specific-abstract		-	-1.47*
	declarative-specific-concrete			-

Note. * denotes a significant result in pairwise comparison, $p < .05$.

For the declarative suggestion statements, there was a significant difference in ratings on all four dependent measures between the three variations: helpfulness ($\chi^2(2) = 44.29, p < .001$), clarity ($\chi^2(2) = 35.75, p < .001$), amount of detail ($\chi^2(2) = 29.37, p < .001$), and confidence ($\chi^2(2) = 45.78, p < .001$). Respondents gave significantly higher ratings on the four measures for the declarative-password-specific-concrete statement than for the other statements. Similar to the suggestions presented at the before-interaction step, the pairwise comparison showed a significant difference between the abstract and concrete types of declarative suggestion on the four dependent measures (see Table 7.18): the concrete suggestion was rated higher than the abstract one. There was also a significant difference between the general and specific abstract declarative suggestions. Interestingly, the former was rated higher than the latter. On the other hand, providing a concrete specific declarative suggestion was perceived to be more helpful and respondents felt more confident reading this type of suggestion than the general one.

For the procedural suggestion statements, there were significant differences in ratings of helpfulness ($\chi^2(5) = 74.96, p < .001$), clarity ($\chi^2(5) = 63.42, p < .001$), amount of detail ($\chi^2(5) = 49.08, p < .001$), and confidence ($\chi^2(5) = 75.25, p < .001$) between the six variations. Respondents gave significantly higher ratings on the four measures for the imperative-specific-concrete statement than for the other statements. Similar to the suggestion statements provided at the before-interaction step, the pairwise comparison showed significant differences between the abstract and concrete procedural suggestions on the four dependent measures (see Table 7.19). There was also a significant difference between the general and specific procedural suggestions, as long as the suggestion was written concretely. Interestingly, the negative format of procedural suggestion was rated significantly higher than the specific abstract procedural suggestion, but significantly lower than the specific concrete procedural suggestion.

Table 7.19 Pairwise comparisons between the procedural suggestion statements at the during-
interaction step across the dependent measures

		imperat ive- general	imperati ve- general- polite	imperativ e-specific- abstract	imperativ e-specific- concrete	imperativ e-specific- abstract- polite	imperativ e- specific- negative
Helpfulness	imperative-general	-	-0.22	0.29	-2.46*	0.49	-0.75
	imperative-general-polite		-	0.52	-2.24*	0.71	-0.53
	imperative-specific-abstract			-	-2.75*	0.19	-1.04*
	imperative-specific-concrete				-	2.94*	1.71*
	imperative-specific-abstract- polite					-	-1.24*
	imperative-specific-negative						-
Clarity	imperative-general	-	-0.22	0.57	-2.10*	0.68	-0.52
	imperative-general-polite		-	0.79	-1.88*	0.90*	-0.29
	imperative-specific-abstract			-	-2.68*	0.10	-1.09*
	imperative-specific-concrete				-	2.78*	1.59*
	imperative-specific-abstract- polite					-	-1.19*
	imperative-specific-negative						-
Amount of detail	imperative-general	-	0.09	-0.24	1.96*	-0.28	0.68
	imperative-general-polite		-	0.32	-1.87*	0.37	-0.59
	imperative-specific-abstract			-	-2.19*	0.04	-0.91*
	imperative-specific-concrete				-	2.24*	1.28*
	imperative-specific-abstract- polite					-	-0.97*
	imperative-specific-negative						-
Confidence	imperative-general	-	-0.22	0.59	-2.27*	0.74	-0.43
	imperative-general-polite		-	0.81	-2.04*	0.96*	-0.21
	imperative-specific-abstract			-	-2.85*	0.15	-1.02*
	imperative-specific-concrete				-	3.00*	1.84*
	imperative-specific-abstract- polite					-	-1.16*
	imperative-specific-negative						-

Note. * denotes a significant result in pairwise comparison, $p < .05$.

7.3.2.2.3 Error Message Instructions

Table 7.20 shows the mean (and median) ratings for the perceived helpfulness, clarity, amount of detail, and respondents' confidence regarding the declarative error message statements provided at the during-interaction step; Table 7.21 presents the results of the pairwise comparison across the dependent measures.

Table 7.20 Mean (median) ratings of the dependent measures: error message at the during-interaction step

	Helpfulness	Clarity	Amount of detail	Confidence
<i>Declarative error message statements (3 variations)</i>				
declarative-general-positive	3.26 (3.00)	3.40 (4.00)	2.30 (2.00)	2.95 (3.00)
declarative-general-negative	2.72 (3.00)	3.00 (3.00)	2.09 (2.00)	2.79 (3.00)
declarative-specific-negative	2.65 (3.00)	2.93 (3.00)	2.05 (2.00)	2.74 (3.00)
<i>p value</i>	.014	n.s.	n.s.	n.s.

Table 7.21 Pairwise comparisons between the declarative error message statements at the during-interaction step for perceived helpfulness

		declarative-general-positive	declarative-general-negative	declarative-specific-negative
Helpfulness	declarative-general-positive	-	0.38	0.45*
	declarative-general-negative		-	0.70
	declarative-specific-negative			-

Note. * denotes a significant result in pairwise comparison, $p < .05$.

There was a significant difference in ratings of perceived helpfulness ($\chi^2(2) = 8.56$, $p = .014$) between the three variations of the declarative error message statements, but not in the ratings of clarity ($\chi^2(2) = 5.35$, $p = .069$), amount of detail ($\chi^2(2) = 4.120$, $p = .127$), or confidence ($\chi^2(2) = 0.438$, $p = .804$). Respondents gave significantly higher ratings on perceived helpfulness for the declarative-general-positive statement than for the other statements. The pairwise comparison showed no significant difference between the positive and negative formats if the error message statement was a general one (see Table 7.21). Interestingly, a positive general error message was perceived as significantly more helpful than a negative specific error message.

The following summarises the results regarding the types of instruction presented at the during-interaction step. For the policy statements, perceived helpfulness, clarity, amount of detail, and users' confidence were higher for the procedural format than for the declarative one. Users preferred a positive imperative policy statement regardless of whether or not it included a politeness element. In contrast, for the suggestion statements, the perceived helpfulness, clarity, amount of detail, and users' confidence were higher for the declarative format than for the procedural one. Users preferred a specific concrete suggestion. Finally, respondents also preferred the error statement to be general and positive.

7.3.2.3 Instructions at the After-Interaction Step

7.3.2.3.1 Policy Instructions

Table 7.22 shows the mean (and median) ratings for the perceived helpfulness, clarity, amount of detail, and respondents' confidence regarding the policy statements provided at the after-interaction step; Table 7.23 and Table 7.24 present the results of the pairwise comparison for the declarative and procedural policy statements across the dependent measures, respectively.

Table 7.22 Mean (median) ratings of the dependent measures: policy at the after-interaction step

	Helpfulness	Clarity	Amount of detail	Confidence
<i>Declarative policy statements vs. procedural policy statements</i>				
Overall declarative	3.43 (3.67)	3.47 (3.67)	2.49 (2.67)	3.44 (3.67)
Overall procedural	3.45 (3.67)	3.53 (3.67)	2.56 (2.67)	3.39 (3.67)
<i>p value</i>	n.s.	n.s.	n.s.	n.s.
<i>Declarative policy statements (3 variations)</i>				
declarative-positive	3.88 (4.00)	3.84 (4.00)	2.56 (3.00)	3.79 (4.00)
modal-positive	3.81 (4.00)	3.84 (4.00)	2.63 (3.00)	3.72 (4.00)
modal-negative	2.60 (3.00)	2.72 (3.00)	2.28 (3.00)	2.81 (3.00)
<i>p value</i>	.000	.000	n.s.	.000
<i>Procedural policy statements (3 variations)</i>				
imperative-positive	3.93 (4.00)	4.02 (4.00)	2.65 (3.00)	3.74 (4.00)
imperative-positive-polite	3.81 (4.00)	3.84 (4.00)	2.70 (3.00)	3.77 (4.00)
imperative-negative	2.60 (3.00)	2.72 (3.00)	2.33 (2.00)	2.65 (3.00)
<i>p value</i>	.000	.000	.001	.000

Overall, there were no significant differences in ratings of helpfulness ($Z = -0.25$, $p = .804$), clarity ($Z = -0.80$, $p = .422$), amount of detail ($Z = -1.44$, $p = .151$), and confidence ($Z = -0.53$, $p = .599$) between the two types of format of policy instructions.

For the declarative policy statements, there were significant differences in ratings of helpfulness ($\chi^2(2) = 42.88$, $p < .001$), clarity ($\chi^2(2) = 26.72$, $p < .001$), and confidence ($\chi^2(2) = 26.43$, $p < .001$) between the three variations, but not in ratings of the amount of detail ($\chi^2(2) = 5.56$, $p = .062$). Respondents gave significantly higher ratings on the three measures for the declarative-positive statement than for the other statements. Furthermore, the pairwise comparison showed a significant difference between the

negative and positive formats (see Table 7.23): the positive one was always rated higher.

Table 7.23 Pairwise comparisons between the declarative policy statements at the after-interaction step across the dependent measures

		declarative-positive	modal-positive	modal-negative
Helpfulness	declarative-positive	-	0.16	1.06*
	modal-positive		-	0.90*
	modal-negative			-
Clarity	declarative-positive	-	0.05	0.79*
	modal-positive		-	0.74*
	modal-negative			-
Confidence	declarative-positive	-	0.12	0.79*
	modal-positive		-	0.67*
	modal-negative			-

Note. * denotes a significant result in pairwise comparison, $p < .05$.

For the procedural policy statements, there were significant differences in ratings on the four dependent measures between the three variations of the procedural policy: helpfulness ($\chi^2(2) = 31.49$, $p < .001$), clarity ($\chi^2(2) = 21.15$, $p < .001$), amount of detail ($\chi^2(2) = 13.57$, $p = .001$), and confidence ($\chi^2(2) = 31.46$, $p < .001$). Respondents gave significantly higher ratings on helpfulness and clarity for the imperative-positive statement, and higher ratings on the amount of detail and confidence for the imperative-positive-polite statement than for the other statements. Similar to the procedural policy statements provided at the before-interaction and during-interaction steps, the pairwise comparison showed no significant differences in the ratings of the four dependent measures between the imperative-positive and imperative-positive-polite statements (see Table 7.24). However, there was a significant difference between the negative and positive formats of procedural policy statements.

Table 7.24 Pairwise comparisons between the procedural policy statements at the after-interaction step across the dependent measures

		imperative- positive	imperative- positive-polite	imperative- negative
Helpfulness	imperative-positive	-	0.05	0.86*
	imperative-positive-polite		-	0.81*
	imperative-negative			-
Clarity	imperative-positive	-	0.09	0.74*
	imperative-positive-polite		-	0.65*
	imperative-negative			-
Amount of detail	imperative-positive	-	-0.70	0.45
	imperative-positive-polite		-	0.45*
	imperative-negative			-
Confidence	imperative-positive	-	-0.15	0.74*
	imperative-positive-polite		-	0.90*
	imperative-negative			-

Note. * denotes a significant result in pairwise comparison, $p < .05$.

7.3.2.3.2 Suggestion Instructions

Table 7.25 shows the mean (and median) ratings for the perceived helpfulness, clarity, amount of detail, and respondents' confidence regarding the suggestion statements provided at the after-interaction step; Table 7.26 presents the results of the pairwise comparison for the procedural suggestion statements across the dependent measures.

Table 7.25 Mean (median) ratings of the dependent measures: suggestion at the after-interaction step

	Helpfulness	Clarity	Amount of detail	Confidence
<i>Declarative suggestion statement (1 variation) vs. procedural suggestion statements</i>				
declarative-specific-concrete	3.16 (3.00)	3.33 (3.00)	2.33 (2.00)	3.37 (3.00)
Overall procedural	2.49 (2.33)	2.61 (2.33)	1.94 (1.67)	2.59 (2.33)
<i>p value</i>	.000	.001	.005	.000
<i>Procedural suggestion statements (3 variations)</i>				
imperative-general	2.05 (2.00)	2.07 (2.00)	1.67 (1.00)	2.16 (2.00)
imperative-general-polite	2.14 (2.00)	2.16 (2.00)	1.58 (1.00)	2.16 (2.00)
imperative-specific	3.28 (3.00)	3.60 (4.00)	2.56 (3.00)	3.44 (3.00)
<i>p value</i>	.000	.000	.001	.000

Overall, there were significant differences in ratings of helpfulness ($Z = -3.58$, $p < .001$), clarity ($Z = -3.41$, $p = .001$), amount of detail ($Z = -2.81$, $p = .005$), and confidence ($Z = -3.97$, $p < .001$) between the two formats of the suggestion

instructions. Similar to the suggestion presented at the before-interaction and after-interaction steps, respondents gave significantly higher ratings on all dependent measures for the declarative suggestion than the procedural one.

Table 7.26 Pairwise comparisons between the procedural suggestion statements at the after-interaction step across the dependent measures

		imperative-general	imperative-general-polite	imperative-specific
Helpfulness	imperative-general	-	-0.70	-0.91*
	imperative-general-polite		-	-0.84*
	imperative-specific			-
Clarity	imperative-general	-	-0.02	-0.92*
	imperative-general-polite		-	-0.90*
	imperative-specific			-
Amount of detail	imperative-general	-	0.12	-0.87*
	imperative-general-polite		-	-0.98*
	imperative-specific			-
Confidence	imperative-general	-	0.01	-0.88*
	imperative-general-polite		-	-0.90*
	imperative-specific			-

Note. * denotes a significant result in pairwise comparison, $p < .05$.

For the procedural suggestion statements, there were significant difference in ratings of helpfulness ($\chi^2(2) = 34.57$, $p < .001$), clarity ($\chi^2(2) = 31.22$, $p < .001$), amount of detail ($\chi^2(2) = 38.32$, $p = .001$), and confidence ($\chi^2(2) = 33.35$, $p < .001$) between the three variations of the procedural suggestion. Respondents gave significantly higher ratings on the four measures for the imperative-specific statement than for the other statements. Furthermore, the pairwise comparison showed significant differences between the general and specific procedural suggestions on the four dependent measures (see Table 7.26). In addition, there were no significant differences between the ratings on the four dependent measures for the statements when a politeness element was added.

7.3.2.3.3 Error Message Instructions

Table 7.27 shows the mean (and median) ratings for the perceived helpfulness, clarity, amount of detail, and respondents' confidence regarding the declarative error message statements provided at the after-interaction step; Table 7.28 presents the results of the pairwise comparison across the dependent measures.

Table 7.27 Mean (median) ratings of the dependent measures: error message at the after-interaction step

	Helpfulness	Clarity	Amount of detail	Confidence
<i>Declarative error message statements (5 variations)</i>				
declarative-general-negative	2.70 (3.00)	2.91 (3.00)	2.14 (2.00)	2.65 (3.00)
declarative-general-negative-polite	1.42 (1.00)	1.49 (1.00)	1.37 (1.00)	1.77 (1.00)
declarative-general-positive	2.77 (2.00)	2.74 (3.00)	2.12 (2.00)	2.70 (3.00)
declarative-specific-negative	3.26 (3.00)	3.44 (3.00)	2.44 (2.00)	3.05 (3.00)
phrasal-general-negative	1.58 (1.00)	1.67 (1.00)	1.37 (1.00)	1.74 (1.00)
<i>p value</i>	.000	.000	.000	.000

Table 7.28 Pairwise comparisons between the declarative error message statements at the after-interaction step across the dependent measures

	declarative-general-negative	declarative-general-negative-polite	declarative-general-positive	declarative-specific-negative	phrasal-general-negative
Helpfulness	declarative-general-negative	-	-1.72*	-0.19	-0.72*
	declarative-general-negative-polite		-	-1.91*	-2.44*
	declarative-general-positive			-	-0.54
	declarative-specific-negative				-
	phrasal-general-negative				
Clarity	declarative-general-negative	-	-1.79*	0.23	-0.52
	declarative-general-negative-polite		-	-1.56*	-2.31*
	declarative-general-positive			-	-0.76*
	declarative-specific-negative				-
	phrasal-general-negative				
Amount of detail	declarative-general-negative	-	-1.34*	-0.04	-0.59
	declarative-general-negative-polite		-	-1.37*	-1.93*
	declarative-general-positive			-	-0.56
	declarative-specific-negative				-
	phrasal-general-negative				
Confidence	declarative-general-negative	-	-1.20*	-0.11	-0.52
	declarative-general-negative-polite		-	-1.30*	-1.72*
	declarative-general-positive			-	-0.42
	declarative-specific-negative				-
	phrasal-general-negative				

Note. * denotes a significant result in pairwise comparison, $p < .05$.

There were significant differences in ratings for the perceived helpfulness ($\chi^2(4) = 100.81$, $p < .001$), clarity ($\chi^2(4) = 87.21$, $p < .001$), amount of detail ($\chi^2(4) = 78.43$, $p < .001$), and confidence ($\chi^2(4) = 64.35$, $p < .001$) between the five variations of the declarative error message statements. Respondents gave significantly higher ratings

on the four dependent measures for the declarative-specific-negative statement than for the other statements. Similar to the error message provided at the during-interaction step, the pairwise comparison showed no significant difference between the positive and negative formats if the error message statement was general (see Table 7.28).

The following summarises the results regarding the types of instruction presented at the after-interaction step. There was no significant difference on the perceived helpfulness, clarity, amount of detail, and users' confidence between declarative and procedural password policy statements. For the creation suggestion statements, users preferred a specific imperative format. Finally, users preferred the error message statement to be specific and negative.

7.3.2.4 Users' Common Password Creation Practices

On average, respondents reported that they had around 35.87 (standard deviation = 66.16) password-protected accounts and approximately 16.97 (standard deviation = 42.15) passwords. However, the majority reported that they used the same password (94, 80.30%) or slightly different passwords (88, 75.20%) for multiple accounts. In terms of their frequency of changing passwords, the majority of respondents (25, 21.40%) changed their passwords every six months, and only 15 (12.80%) of them never changed their passwords.

Respondents were asked to indicate their knowledge of what makes a secure password, and the majority (49, 41.90%) described themselves as very knowledgeable. Moreover, they had a similar perception of what makes a secure password as those who participated in Study 4 (see Section 6.3.3, Chapter 6). Most respondents commented that a secure password should have a combination of different character classes and should not be based on dictionary words or common phrases. However, a few respondents also mentioned length and changing passwords frequently as criteria for secure passwords. Over a third of respondents (43, 36.75%) were very confident about the strength of their most complicated password, whereas nearly 10% of them (11, 9.40%) felt not at all confident.

Regarding the password creation instructions, over a third of respondents (32, 27.35%) indicated that they read instructions when creating a new password ‘always’. However, the following were circumstances in which they did not do so: if instructions were too lengthy or invisible, if the password was for a low-value account, if they lacked time, if they managed to establish a password on their first attempt, if they were familiar with the instructions on the website (e.g. changing current passwords), and finally, if they used a password manager. Again, these circumstances were highly similar to the ones found in Study 4 (see Section 6.3.3, Chapter 6).

7.4 Discussion

This study had two aims: to understand what kinds of instructions are provided by PCSs to support users in creating passwords, and to investigate users’ perceptions of the most frequently used instructions. To address these aims, the study analysed a total of 95 instructions extracted from 27 PCSs. Based on this analysis, an online questionnaire study was conducted with 117 respondents to understand how the most frequently used instructions affect users’ perceptions of the instructions.

The first research question (RQ1) concerned the types of user instructions found in current PCSs, while the second (RQ2) related to the instruction format, declarative or procedural, used in writing each type of instruction. The third question (RQ3) examined the difference between the declarative and procedural formats for each type of instruction in terms of perceived helpfulness, clarity, amount of detail, and users’ confidence. Finally, the fourth question (RQ4) compared the commonly used instructions to the users’ preference. These research questions are answered according to the timings of presentation, as follows: instruction presentation at the before-interaction step in Section 7.4.1, at the during-interaction step in Section 7.4.2, and at the after-interaction step in Section 7.4.3.

7.4.1 Instructions at the Before-Interaction Step

The results revealed that only 10.52% of instructions are provided before users start interacting with PCSs. Two types of user instruction were identified at this step:

password policy and password creation suggestions. Almost all user instructions at this step were about password policy (90.00%); only one instance of password creation suggestion (10.00%) was found.

For the policy instruction, the use of the declarative format was more common than the procedural one in current PCSs. Furthermore, within the declarative format, almost all policy statements were written using a phrasal-positive construction (see Table 7.2). These results are consistent with the users' preferences to some extent. In general, users perceived the declarative policy to be clearer and more helpful. They also felt more confident reading declarative policy statements than procedural ones. However, within the declarative-format statements, the phrasal-positive construction was rated lower than the declarative-positive construction in terms of the instruction's helpfulness and the users' confidence (see Table 7.9). Users commented that the use of the phrasal-positive construction made the policy less user-friendly, less clear, and more forceful and abrupt. For example, one user wrote: *'When the instructions are embedded within a single sentence the meaning seems clearer to me whereas [the phrasal-positive construction] introduces a level of uncertainty'*. Hence, it can be assumed that the current practice of using declarative policy statements before interaction with the PCS matches users' preferences. However, the construction of the declarative policy statements does not meet users' needs: the user study indicates a preference for the declarative-positive construction of declarative policy statements, since it makes policy instructions easy to parse and understand with little room for misinterpretation.

Only one instance of suggestion instruction was found in the examined PCSs. This statement was written using the declarative format with an action-oriented style (see Table 7.2). Overall, the participants in the user study gave significantly higher ratings on helpfulness, clarity, amount of detail, and their confidence when reading declarative suggestion statements than procedural ones. Within the declarative statements, the concrete construction was always rated significantly higher on the four dependent measures than the abstract construction, regardless of whether the suggestion was written using a password-oriented or an action-oriented style (see Table 7.11). Participants in the user study commented that the use of the abstract

construction made the suggestion somewhat vague and unhelpful. For example, one user wrote: *'it's appealing to those who want to create a good password, but not clear enough'*. Thus, it seems that the current practice of using declarative suggestions before interaction with the PCS matches user preferences. Furthermore, the user study suggests the use of a concrete (password- or action-oriented) construction for the declarative suggestion, since it makes the suggestion concise, helpful, and straightforward. However, with the lack of implementation of suggestion instructions in current PCSs at this stage, caution is needed, as the findings regarding current practices might not be generalisable.

As a final note on the instructions found in current PCSs at the before-interaction step, it is somewhat surprising that there were so frequently no instructions for the users before they started creating a password. This could explain why users tend to make poor password choices. Current PCSs do not provide enough support up front in terms of user instructions on how to create passwords, which could have an influence on users' performance.

7.4.2 Instructions at the During-Interaction Step

Nearly half (47.37%) of the analysed instructions were presented during the password entry stage. Three types of instruction were identified: password policy, password creation suggestions, and error messages. Most user instructions concerned password policy (55.56%), followed by password creation suggestions (28.89%) and error messages (15.56%).

For the policy instruction, the use of the declarative format was more common than the procedural one in current PCSs. Within the declarative format, there was an even distribution of policy instructions between the declarative constructions (i.e. modal, phrasal, and declarative) (see Table 7.3). The results of the analysis do not align with the users' preferences. In general, the user study showed that participants gave significantly higher ratings on helpfulness, clarity, amount of detail, and their confidence when reading procedural policy statements than declarative ones. Furthermore, among the procedural statements, users preferred the imperative-positive

construction (with/without politeness element) to the imperative-negative construction during their interaction with the PCSs (see Table 7.9). Users commented that the use of the imperative-negative construction made the policy less user-friendly and less clear. For example, one participant wrote: *'Positive rules are clearer than negatives'*. It can therefore be assumed that the current practice of using declarative policy during interaction with the PCS does not match users' preferences. Instead, the user study suggests the use of the imperative-positive construction for procedural policy statements, since it makes instructions clear and precise. In addition, users found that adding a politeness element to procedural policy statements made them highly user-friendly.

For the suggestion instructions, the use of the procedural format was more common than the declarative one in current PCSs (see Table 7.3). Furthermore, within the procedural format, the use of specific and negative constructions was more common than the use of general and positive constructions. Again, the results of the analysis do not align with the users' preferences regarding the suggestion instructions. In general, the user study showed that the respondents gave higher ratings on helpfulness, clarity, amount of detail, and their confidence when reading declarative suggestion statements than procedural ones. Furthermore, users preferred a specific concrete construction for the declarative suggestions during their interaction with the PCSs (see Table 7.17). They commented that the use of the general and abstract construction made the declarative suggestion vague and unhelpful. For example, one respondent wrote that the general declarative construction *'doesn't explain what it wants'*. Thus, it seems that in general, the current practice of using procedural suggestions during interaction with the PCS does not match users' preferences. Instead, the user study suggests the use of the specific concrete construction since it makes the suggestion instructions clear, informative, easy to understand, and straightforward.

Regarding error message instructions, all were written using a declarative format, and the use of general and negative constructions was common (see Table 7.3). The results of the analysis are consistent with the users' preferences to some extent. In general, users perceived the declarative error message to be more helpful when it was written using a general and positive construction (see Table 7.20). In contrast, users

commented that the use of the negative error message construction during password entry made the creation process annoying and might distract them. For example, one respondent wrote: *'As I am typing, it is not helpful, wait until I have finished. This only serves to distract me, and it is hard enough already!'*. Thus, the results suggest that the current practice of using general declarative error messages during password entry matches users' preferences to some extent. However, the negative construction of statements does not meet user needs: the present results suggest that using a positive general construction for declarative error messages is preferable, since it makes the error message more helpful.

7.4.3 Instructions at the After-Interaction Step

The results revealed that about 40.12% of the instructions analysed were presented after the password was entered in full. The instruction types presented at this stage were similar to those presented at the during-interaction step. Most user instructions consisted of password policy (47.50%), followed by password creation suggestions (27.50%) and error messages (25.00%).

For the policy instructions, similar to the two other timings of interaction, the use of the declarative format was generally more common than the procedural one in current PCSs. Within the declarative format, the modal construction was more common than the declarative one (see Table 7.4). Contrary to expectations, the findings from the user study showed no significant difference between the declarative and procedural policy statements at this stage. However, the results did show that users preferred both declarative and modal constructions for the declarative format, and the imperative-positive (with/without politeness element) construction for the procedural format (see Table 7.22). Thus, it can be assumed that, for the declarative format, the current practice of using modal policy statements after interaction with the PCS matches users' preferences.

For the suggestion instructions, similar to the during-interaction step, the use of the procedural format was overall frequent in current PCSs, with no occurrence of declarative suggestions at this stage. Furthermore, the general construction of

procedural suggestions was much more common than the specific one (see Table 7.4). However, the results of the user study do not align with the current practice of using the procedural format after interaction with the PCS. Similar to the during-interaction step, users preferred a specific concrete construction for the declarative suggestions after their interaction with the PCSs (see Table 7.25). The users indicated that they found this construction helpful and straightforward. However, it is important to note that the imperative-specific construction of the procedural suggestion did not seem to differ from the specific concrete construction of the declarative suggestion (based on the mean ratings in Table 7.25). Thus, this interpretation must be used with caution, because there was only one variation of the declarative format.

Similar to the during-interaction step, all error message instructions were written using a declarative format, and the use of the declarative, general, and negative constructions was common (see Table 7.4). The results of the analysis seem to be consistent to some extent with the users' preferences. In general, the participants perceived the declarative error message to be clearer and more helpful when it was written using a specific and negative construction; they also felt more confident reading this type of construction (see Table 7.27). In contrast, the users commented that the general error message construction at this stage was less helpful and too vague. For example, one respondent wrote: *'It only broadly explains the problem'*. It can thus be stated that the current practice of using negative declarative error messages after password entry somewhat matches users' preferences. However, the construction of general statements is not in line with user needs. Instead, the user study suggests the use of the specific negative construction for declarative error messages since it makes the message more helpful at this stage.

7.5 Conclusions

All in all, these findings provide additional evidence that users struggle to understand the instructions to create good and secure passwords. These results are in line with those of the other studies (see Studies 2 and 3, Chapters 4 and 5, respectively) presented in this thesis, which showed that most of the usability problems with user instructions were related to the instructions' amount of detail and clarity. Therefore,

the combined qualitative (Studies 2 and 3) and quantitative (Study 5) data may help us understand why users have difficulty choosing secure passwords, since user instructions play a key role in understanding password requirements. However, these findings must be interpreted with caution because they are based on self-reported data in an artificial password creation situation. Thus, further research is needed to better understand the effects of the user instructions on the quality of passwords by asking users to create password while reading different sets of instructions.

Table 7.29 Summary of the current practices and user study findings

	before-interaction step	during-interaction step	after-interaction step
<i>Password policy</i>			
Current practice	Declarative (phrasal-positive)	Declarative (modal-positive)	Declarative (modal-positive)
User study	Declarative (declarative-positive)	Procedural (imperative-positive-with(out)-polite)	Declarative and procedural ([declarative/modal]-positive) (imperative-positive-with(out)-polite)
<i>Password creation suggestion</i>			
Current practice	Declarative (concrete-action-oriented)	Procedural (imperative-specific-negative)	Procedural (imperative-general)
User study	Declarative (concrete-[action/password]-oriented)	Declarative (declarative-concrete-specific)	Declarative (declarative-concrete-specific)
<i>Error message</i>			
Current practice		Declarative (declarative-general-negative)	Declarative (declarative-general-negative)
User study		Declarative (declarative-general-positive)	Declarative (declarative-specific-negative)

To conclude, one of the main findings to emerge from this study is that current practices of user instructions vary widely and do not match users' needs, as shown in Table 7.29. In general, the user study suggests the use of declarative policy statements before users start interacting with PCs; and the use of procedural policy statements while and after the users enter their password. The results also indicate the benefit of using declarative suggestions through the password creation process regardless of the timing of presentation. Finally, users appear to prefer the use of positive general declarative

error messages during the password entry stage, and the use of negative specific declarative error messages after the password is entered in full.

Chapter 8

The Individual Effects of Supporting Features on Password Creation and Recall – *Study 6*

8.1 Introduction

Most supporting features currently available in PCSs appear to be implemented in an ad hoc manner instead of by examining users' needs (while always considering security issues): for instance, by investigating what supporting features users want to have, and when and how users want particular features to be presented. The exploratory analysis of current PCSs (see Study 1, Chapter 3) revealed that current implementations of supporting features are very inconsistent. This causes usability problems and affects the quality of passwords, in addition to potentially confusing users as they move from one PCS to another. Moreover, the results from the usability evaluations (see Table 5.9, Chapter 5) confirmed the need to improve the design and usability of supporting features. A number of usability problems related to the password policies (24.79% of all distinct usability problems reported by either experts and users), creation suggestions (16.52%), and strength indicators (9.91%). In general, some of these problems were related to the lack of supporting features, whereas others were related to the presentation of these features. Therefore, it is clearly important to examine the individual effects of presenting the supporting features in PCSs and to ensure that they support users well when creating passwords.

Evidence from the literature shows the importance of providing guidance during the password creation process to improve the quality of passwords. For example, a study by Furnell and Bär (2013) shows that even the basic presence of guidance without

any compliance enforcement has a positive effect on the quality of the passwords created. A number of studies have investigated different ways in which PCSs encourage appropriate user behaviour in creating passwords, such as the use of password strength indicators. Among them, Ur et al. (2012) evaluated the effectiveness of password strength indicators using 14 different configurations of the indicators. They found that indicators that scored passwords stringently caused users to create passwords that were longer and that contained more uppercase letters and non-alphabetic elements. In contrast, in two studies that they argued were more realistic, Egelman et al. (2013) found that password strength indicators only influenced password strength when the password was associated with a high-risk account. Thus, password creation behaviour is dependent on the context in which the password is to be used. However, password strength indicators are only one of a range of features that occur in current PCSs to encourage appropriate user behaviour. Other features include statements of password policy, creation suggestions for strong passwords, and tips on how to create passwords.

Therefore, the present study aims to understand the individual effect of presenting the password policy, creation suggestions, and strength indicator to users in a PCS. The study specifically examines the best timing at which to present the password policy and creation suggestion. It also investigates the best media and colour-scheme presentation for the strength indicators. To address this aim, an online non-factorial mixed design study was conducted. The study consisted of two parts: password creation (Part I) and password recall (Part II). In Part I, each participant was asked to create a number of passwords using different supporting features. In Part II, all participants were asked to recall their passwords three days later. A total of 257 participants from MTurk completed Part I although only 168 participants (65.36%) returned to Part II.

The following research questions are formulated to address the aims of this study.

RQ1. Are there differences in PCS usability and password strength between different timings of presentation of password policy when users create passwords?

RQ2. Are there differences in PCS usability between different timings of presentation of password policy when users recall passwords?

RQ3. Are there differences in PCS usability and password strength between different timings of presentation of password creation suggestions when users create passwords?

RQ4. Are there differences in PCS usability between different timings of presentation of password creation suggestions when users recall passwords?

RQ5. Are there differences in PCS usability and password strength between different media and colour-scheme presentations of password strength indicators when users create passwords?

RQ6. Are there differences in PCS usability between different media and colour-scheme presentations of password strength indicators when users recall passwords?

RQ7. Are there differences in PCS usability and password strength between providing supporting features and not providing them (i.e. baseline) when users create passwords?

RQ8. Are there differences in PCS usability between providing supporting features and not providing them (i.e. baseline) when users recall passwords?

8.2 Method

8.2.1 Design

Table 8.1 Study design and conditions

	type of supporting feature**				
	baseline	policy	suggestion	indicator	
			colour-scheme**		
		timing of presentation*	timing of presentation*	3colour media*	single-colour media*
Group 1	baseline	policy-before-interaction	suggestion-before-interaction	3colour-graphical	single-graphical
Group 2	baseline	policy-during-interaction	suggestion-during-interaction	3colour-textual	single-textual
Group 3	baseline	policy-after-interaction	suggestion-after-interaction	3colour-graphical&textual	single-graphical&textual
Group 4	baseline	policy-during&after-interaction	suggestion-during&after-interaction		

Note. * denotes a between-participants independent variable, ** denotes a within-participants independent variable

Table 8.1 illustrates the study design and conditions. This study used a non-factorial mixed design with two between-participants factors and two within-participants factors.

The first between-participants factor is the timing of presentation for the policy and suggestion statements with four conditions: *before-interaction*, *during-interaction*, *after-interaction*, and *during&after-interaction*. The four conditions of this factor were identified considering the three-step model, as discussed in Study 1 (see Section 3.3.2.1, Chapter 3). The fourth condition, *during&after-interaction*, was included as it was observed from the analysis of current PCSs in Study 1. However, other timings of presentation were also observed but were not included in the study, such as before and during interaction; and before and after interaction. The reason for not including these (apart from the need to not have too many conditions) was that both presentations present the statement before users start interacting and leave it available during the password creation process (i.e. before interaction), and the same statement is then presented again as users interact (i.e. during interaction) or after users interact (i.e. after interaction) with the system. As a result, users see redundant information on the screen using two intersected timings of presentation, which could cause confusion. The second between-participants factor is the type of media presentation used in the strength indicator with three conditions: *graphical*, *textual*, and *graphical&textual*.

Regarding the within-participants factors, the first is the type of supporting feature with four conditions: no supporting feature (*baseline*), password policy (*policy*), password creation suggestion (*suggestion*), and password strength indicator (*indicator*). The second factor is the colour-scheme of the strength indicator with two conditions: *3colour* (using the traffic light metaphor of green/amber/red) and *single-colour*.

This study consisted of two parts. In part I, participants were asked to imagine that they were creating a password for their online bank account. Each participant was asked to create five passwords, one in the baseline condition, two in one of the four timing of presentation conditions, and two in one of the three media conditions. There were thus four groups of participants, one for each of the four timing of presentation

conditions: *before-interaction*, *during-interaction*, *after-interaction* and *during&after-interaction*. Participants were randomly assigned to one of the four groups and exposed to the following conditions along with the baseline condition:

- **Group 1:** participants in this group were exposed to the *before-interaction* condition in both *policy* and *suggestion*; and to the *graphical* condition for the *indicator*.
- **Group 2:** participants in this group were exposed to the *during-interaction* condition in both *policy* and *suggestion*; and to the *textual* condition for the *indicator*.
- **Group 3:** participants in this group were exposed to the *after-interaction* condition in both *policy* and *suggestion*; and to the *graphical&textual* condition for the *indicator*.
- **Group 4:** participants in this group were exposed to the *during&after-interaction* condition in both *policy* and *suggestion*; and randomly assigned to one of three media conditions for the *indicator*. In this manner, a subgroup was assigned to *graphical*, another to *textual*, and a third to *graphical&textual* *indicator*.

In Part II, all participants were asked to recall their five passwords three days later, prompted by the PCS with which they had originally created the password. The design of the different conditions is presented in the Materials in Section 8.2.3.

The dependent measures in the present study were similar to those in Study 4 (see Section 6.2.1, Chapter 6). There were two groups of dependent measures in Part I: those related to the usability of the PCS and those related to the strength of the password.

To investigate the usability of the PCS, two groups of measures were used: efficiency and user satisfaction. The efficiency included two measures: (1) time to create, confirm, and submit the password; and (2) the number of keystrokes used to create a password. The user satisfaction included six measures: participants' ratings (using a 5-point Likert scale: the higher the better) of (1) ease of use, (2) annoyingness, (3)

helpfulness, (4) clarity, (5) amount of detail, and (6) their confidence in using the PCS. Participants were also offered the chance to explain their ratings about the PCS in an optional open-ended question.

Two measures were used to measure password strength: password characteristics and password guessability. The password characteristics included six main measures: (1) password length, (2) number of digits, (3) number of uppercase letters, (4) number of lowercase letters, (5) number of symbols, and (6) number of character classes used in the password (i.e. digits, uppercase letters, lowercase letters, and symbols). Furthermore, additional measures related to particular supporting features: for example, a policy compliance measure for the password policy, suggestion compliance and symbols provision measures for the password creation suggestion, and the password strength score (based on Egelman et al., 2013) for the password strength indicator. The policy and suggestion compliance measures determined whether the passwords complied with the given statements of policy and suggestion. The suggestion provision measure checked whether the passwords included a symbol mentioned in the suggestion statement. Finally, the password guessability included one measure, which was ability to guess the password across five cracking approaches (based on Ur et al., 2015).

Then, in Part II, three dependent measures were used to measure PCS usability: (1) time to recall a password as an efficiency measure, (2) the accuracy of recalling a password as an effectiveness measure, and (3) participants' confidence in recalling the password correctly as the user satisfaction measure.

8.2.2 Participants

A total of 563 participants from MTurk took part in this study; however, 306 entries were excluded because their responses were incomplete (115 entries) or their responses included two or more identical passwords for the different conditions (191 entries). This left a total of 257 participants in the analysis. The criteria of exclusion have been applied carefully to improve the data quality. For more details on the handling and justification of data exclusion, see Section 11.3 in Chapter 11. This included participants randomly assigned to one of the four conditions: *Group 1* (68

participants), *Group 2* (63 participants), *Group 3* (61 participants), and *Group 4* (65 participants). Compensation was provided in the form of USD 0.70 (GBP 0.53) for completing Part I, and a USD 0.70 bonus payment for returning and completing Part II. Table 8.2 summarises the demographic characteristics of the participants in each group and overall.

Table 8.2 Demographic characteristics (frequency and %) of participants in each group and overall

Characteristics		Groups of participants				
		Group 1 (N=68)	Group 2 (N=63)	Group 3 (N=61)	Group 4 (N=65)	Overall (N=257)
Gender	<i>Female</i>	31 (45.59)	29 (46.03)	27 (44.26)	30 (46.15)	117 (45.53)
	<i>Male</i>	37 (54.41)	34 (53.97)	34 (55.74)	35 (53.85)	140 (54.47)
Language	<i>English</i>	62 (91.18)	46 (73.02)	53 (86.89)	56 (86.15)	217 (84.44)
	<i>Other</i>	6 (8.82)	17 (26.98)	8 (13.11)	9 (13.85)	40 (15.56)
Education	<i>School</i>	7 (10.29)	5 (7.94)	8 (13.11)	5 (7.69)	25 (9.73)
	<i>Diploma</i>	16 (23.53)	12 (19.05)	7 (11.48)	14 (21.54)	49 (19.07)
	<i>Bachelor's</i>	26 (38.24)	34 (53.97)	31 (50.82)	38 (58.45)	129 (50.19)
	<i>Master's</i>	16 (23.53)	12 (19.05)	14 (22.95)	8 (12.31)	50 (19.46)
	<i>Doctoral</i>	3 (4.41)	-	1 (1.64)	-	4 (1.57)
Major /Career	<i>Computing</i>	20 (29.41)	24 (38.10)	24 (39.34)	19 (29.23)	87 (33.85)
	<i>Non-computing</i>	48 (70.59)	39 (61.90)	37 (60.66)	46 (70.77)	170 (66.15)

A total of 117 (45.53%) females and 140 (54.47%) males were included in the data analysed. Participants' ages ranged from 18 to 87 years with an average of 34.27 years (standard deviation = 11.18). The native language of 217 (84.44%) participants was English, while the remaining had been speaking English for an average of 21.19 years (standard deviation = 10.44). Half of the participants (129, 50.19%) had a bachelor's degree. The level of education of the remaining participants ranged from postgraduate degree (54, 21.01%) to high school degree (25, 9.73%). Most of the participants (166, 66.1%) were not studying or working in computing fields. On average, the majority of participants spent more than 6 hours a day online and using computers. As shown in Table 8.2, the overall participant characteristics were almost the same between the four groups.

8.2.3 Materials

Two web-based applications were developed: the password creation application used in Part I, and the password recall application used in Part II. The overall design and structure of the two applications were similar to the ones developed in Study 4 (see Section 6.2.3, Chapter 6). The only difference was the number and structure of the password creation pages presented the PCS. This section discusses the design and structure of the two applications in detail.

8.2.3.1 Password Creation Application

Figure 8.1 illustrates the overall structure of the password creation application. The application started with the homepage, followed by the scenario pages.

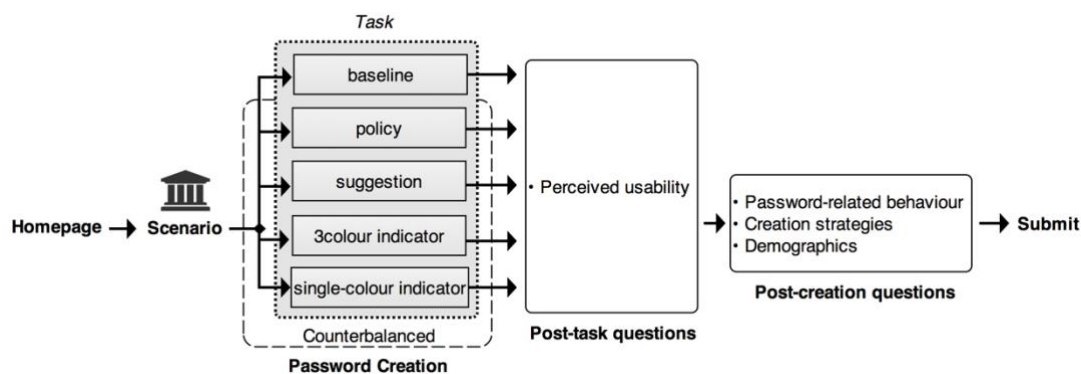


Figure 8.1 Structure of the password creation application used in the creation part

The same online bank account scenario was used in this study as in Study 4, which involved imagining a situation where the participants' online bank account had been compromised and they needed to create a new password using the PCS provided in the study.

As shown in Figure 8.1, five password creation pages were made: *baseline*, *policy*, *suggestion*, *3colour indicator*, and *single-colour indicator*. When the participants successfully confirmed a password (i.e. entered the same password in both fields), an acknowledgement message popped up in each PCS they encountered. After every password creation page, a post-task page appeared that asked participants to rate the PCS. Then, upon completion of the password creation application, a post-creation page appeared to collect information about participants. The following sections

discuss the design and content of password creation page (Section 8.2.3.1.1), post-task page (Section 8.2.3.1.2), and post-creation page (Section 8.2.3.1.3).

8.2.3.1.1 Password Creation Page

The overall design remained the same for the five password creation pages. Figure 8.2 shows an example of one of the five pages, the *baseline* page.

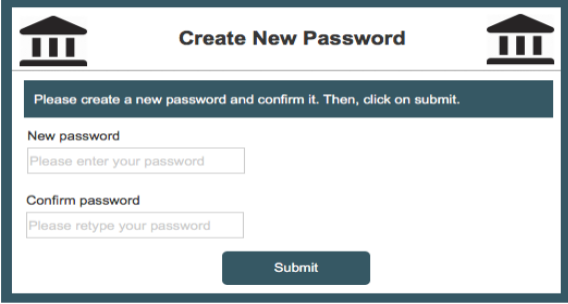


Figure 8.2 A screenshot of the *baseline* password creation page

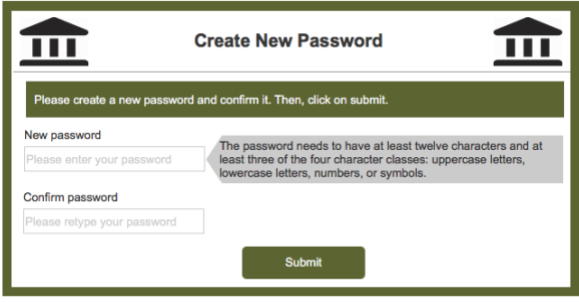
The password creation page contained two fields, one to enter the password and one to confirm it. An additional element varied depending on the type of supporting feature conditions: a statement of password policy for the *policy* page, a statement of creation suggestions for the *suggestion* page, and a password strength indicator for the *indicator* page.

For the *baseline* condition, no supporting feature was presented on the page to help participants create a new password. The *baseline* page was always the first one to be presented in the application among the password creation pages, as shown in Figure 8.1.

For the *policy* condition, a password policy statement was given to participants as a means of helping them create a new password. There were four versions of the *policy* page, one for each of the four timing of presentation conditions. The chosen policy in the present study was the same as the one used in Study 4 (see Section 6.2.3.1.1, Chapter 6): the password had to contain at least 12 characters and at least three of the four character classes – uppercase letters, lowercase letters, numbers, and symbols (based on Shay et al., (2014)). Furthermore, the phrasing of the policy statements was

based on the findings of Study 5 (see Section 7.3.2, Chapter 7) across the different timings of presentation. The four versions of the *policy* page are detailed below:

- *policy-before-interaction* page (see Figure 8.3): The policy was stated using a positive declarative format, as follows: ‘*The password needs to have at least twelve characters and at least three of the four character classes: uppercase letters, lowercase letters, numbers, or symbols.*’ This policy statement was presented before the participants started creating a password, so when they opened the page with the password entry field. The statement remained on the page until the participants submitted their passwords.



The screenshot shows a web form titled "Create New Password" with a header containing two building icons. Below the header is a green instruction bar: "Please create a new password and confirm it. Then, click on submit." The form has two input fields: "New password" with the placeholder "Please enter your password" and "Confirm password" with the placeholder "Please retype your password". A grey tooltip box is positioned over the "New password" field, containing the text: "The password needs to have at least twelve characters and at least three of the four character classes: uppercase letters, lowercase letters, numbers, or symbols." A green "Submit" button is located at the bottom center of the form.

Figure 8.3 A screenshot of the *policy-before-interaction* page

- *policy-during-interaction* page (see Figure 8.4): The policy was stated using a positive imperative format with a politeness element, as follows: ‘*Please use at least twelve characters and at least three of the four character classes: uppercase letters, lowercase letters, numbers, or symbols.*’ This policy statement was presented as the participants started entering a password, so when they put their cursor in the password entry field. The statement disappeared once the participants moved their cursor away from this field and placed it in the confirm password field.

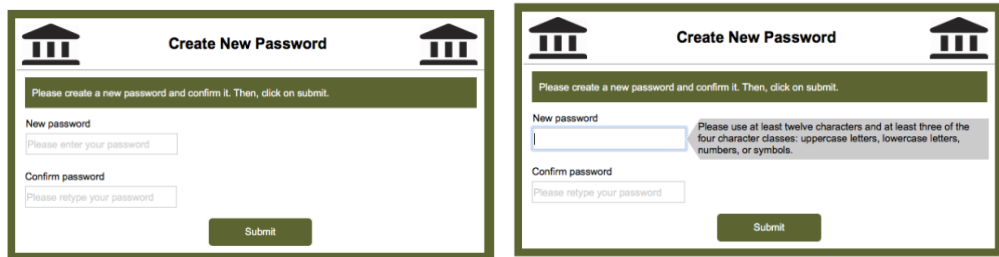


Figure 8.4 Screenshots of the *policy-during-interaction* page

(*left*: before users interacted with password entry field; *right*: while users interacted with password entry field)

- *policy-after-interaction* page (see Figure 8.5): The same policy statement used in the *policy-before-interaction* condition was used on this page, but it was presented after the participants entered the new password in full, so when they put their cursor in the confirm password field. The statement disappeared once the participants moved their cursor away from the confirm password field.

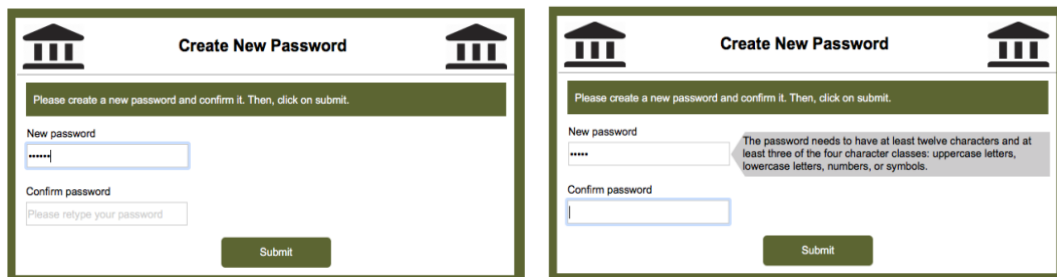


Figure 8.5 Screenshots of the *policy-after-interaction* page

(*left*: while users interacted with password entry field; *right*: after users interacted with password entry field)

- *policy-during&after-interaction* page (see Figure 8.6): This page was the combination of the *policy-during-interaction* and *policy-after-interaction* conditions.

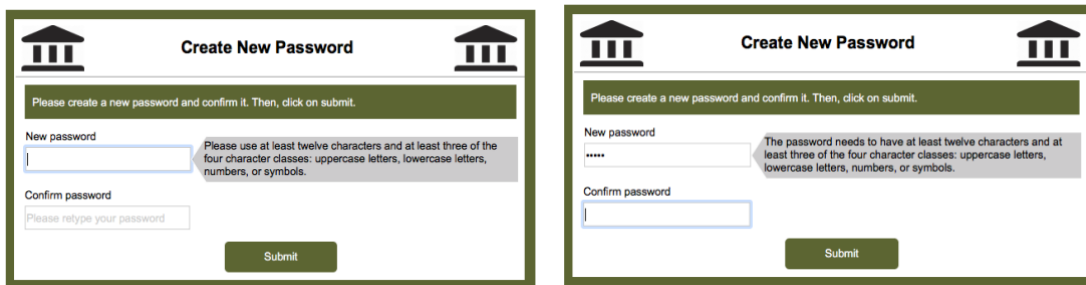


Figure 8.6 Screenshots of the *policy-during&after-interaction* page

(left: while users interacted with password entry field; right: after users interacted with password entry field)

For the *suggestion* condition, a password creation statement was offered in the PCS. Similar to the policy condition, four versions of the *suggestion* page were developed, one for each of the four timing of presentation conditions. In the literature, to the best of the author's knowledge, no recommendations are made about the creation suggestion that should be followed during the password creation process. Therefore, the chosen suggestion was selected considering what was observed in the analysis of current PCSs in Study 1. Furthermore, it was interesting to include a particular set of symbols in the suggestion statement as examples and to see whether their presence influenced participants to use them in their passwords. The chosen symbols were '!', '@', '#', '{', ',' and '~'; the former three were common, whereas the remaining were not. Similar to the policy statements, the findings of Study 5 (see Section 7.3.2, Chapter 7) were used to phrase the suggestion statement across the different timings of presentation. The four versions of the *suggestion* page are detailed below:

- *suggestion-before-interaction* page (see Figure 8.7): The suggestion was stated using a declarative sentence with a concrete action-oriented format, as follows: *'It will be safer if you use a combination of numbers, letters and symbols like ! @ # { ; ~'*. This suggestion statement was presented in the same way as the policy statement in the *policy-before-interaction* page.

The screenshot shows a web form titled "Create New Password" with a header containing two building icons. Below the header is a instruction bar: "Please create a new password and confirm it. Then, click on submit." The form has two input fields: "New password" (with placeholder "Please enter your password") and "Confirm password" (with placeholder "Please retype your password"). A "Submit" button is at the bottom. A grey suggestion box is positioned to the right of the "New password" field, containing the text: "It will be safer if you use a combination of numbers, letters and symbols like ! @ # { ; ~".

Figure 8.7 A screenshot of the *suggestion-before-interaction* page

- *suggestion-during-interaction* page (see Figure 8.8): The suggestion was stated using a declarative sentence with a concrete specific format, as follows: ‘*You can improve your password by having a combination of numbers, letters and symbols like ! @ # { ; ~*’. This suggestion statement was presented in the same way as the policy statement in the *policy-during-interaction* page.

The figure shows two side-by-side screenshots of the "Create New Password" page. The left screenshot shows the form before interaction, with the "New password" field empty. The right screenshot shows the form while the user is interacting with the "New password" field. In this state, a grey suggestion box is positioned to the right of the input field, containing the text: "You can improve your password by having a combination of numbers, letters and symbols like ! @ # { ; ~".

Figure 8.8 Screenshots of the *suggestion-during-interaction* page

(*left*: before users interacted with password entry field; *right*: while users interacted with password entry field)

- *suggestion-after-interaction* page (see Figure 8.9): The suggestion was stated using a specific imperative format, as follows: ‘*Try one with a combination of numbers, letters and symbols like ! @ # { ; ~*’. This suggestion statement was presented in the same way as the policy statement in the *policy-after-interaction* page.



Figure 8.9 Screenshots of the *suggestion-after-interaction* page

(left: while users interacted with password entry field; right: after users interacted with password entry field)

- *suggestion-during&after-interaction* page (see Figure 8.10): This page was the union of the *suggestion-during-interaction* and *suggestion-after-interaction* conditions.

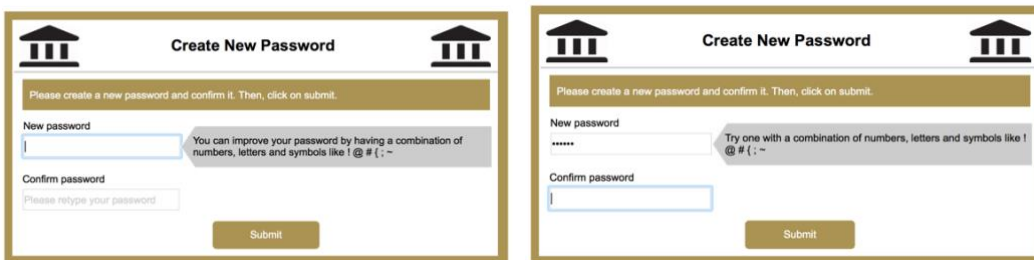


Figure 8.10 Screenshots of the *suggestion-during&after-interaction* page

(left: while users interacted with password entry field; right: after users interacted with password entry field)

For the *indicator* condition, a password strength indicator was provided in the PCS to help users during the password creation process, as shown in see Figure 8.11.

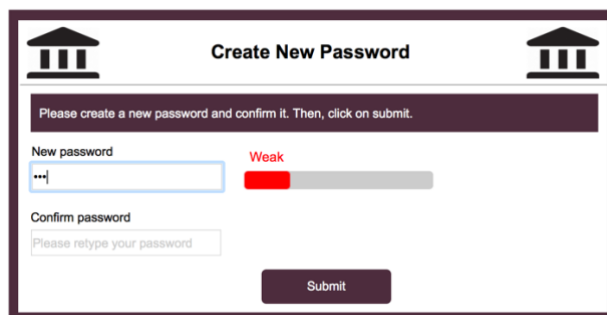


























Figure 8.11 A screenshot of the *3colour indicator (3colour-graphical&textual)* page

A total of six versions of the *indicator* page were created, three for each colour-scheme condition (*3colour* and *single-colour*). Table 8.3 outlines the six password strength indicator conditions.

The *3colour* condition contained three media conditions: *3colour-graphical*, *3colour-textual*, and *3colour-graphical&textual*. For the *single-colour* condition, three media conditions were included: *single-graphical*, *single-textual*, and *single-graphical&textual*. To determine the strength of the password, the same scoring algorithm was used as in Study 4 (see Section 6.2.3.1.1, Chapter 6), which is based on Egelman et al.’s (2013) equation.

Table 8.3 Password strength indicator conditions

		colour-scheme					
media		<i>3colour</i>			<i>single-colour</i>		
<i>graphical</i>	<i>3colour-graphical</i>						
	<i>single-graphical</i>						
<i>textual</i>	<i>3colour-textual</i>	Weak	Medium	Strong	Weak	Medium	Strong
	<i>single-textual</i>	Weak	Medium	Strong	Weak	Medium	Strong
<i>graphical&textual</i>	<i>3colour-graphical&textual</i>						
	<i>single-graphical&textual</i>						

8.2.3.1.2 Post-Task Questions Page

A post-task questions page was provided at the end of the password creation tasks. The page contained questions about the user satisfaction with the PCS; these were the same as those used in Study 4 (see Section 6.2.3.1.2, Chapter 6). This consisted of six measures: (1) ease of use, (2) annoyingness, (3) helpfulness, (4) clarity, (5) amount of detail, and (6) participants’ confidence in using the PCS. These variables were measured using a 5-point Likert item ranging from 1 to 5 where the higher the score, the better.

8.2.3.1.3 Post-Creation Questions Page

A post-creation questions page was provided upon the completion of the password creation application. This page was similar to the one provided in Study 4 (see Section 6.2.3.1.3, Chapter 6). The page listed questions about participants' password-related behaviours, the password creation strategies they used in this study, and information about their demographic characteristics.

8.2.3.2 Password Recall Application

A recall page was developed, the same as the one in Study 4 (see Section 6.2.3.2, Chapter 6), for every password creation page (*baseline*, *policy*, *suggestion*, *3colour indicator*, and *single-colour indicator*) in the password recall application. Each recall page consisted of the task instructions, a screenshot of the PCS used to create the password, a password entry field, and a question about participants' confidence in recalling the correct password. At the end of Part II, a post-recall questions page was presented asking about participants' method of remembering their created passwords and their password management strategies.

8.2.4 Pilot of the Study Procedure

To test the study process and design, a pilot study was conducted, and four PhD students from the Computer Science Department participated. The study procedure was perceived as being smooth. However, two issues were encountered with the task instructions and following them correctly. The first issue concerned the instruction statement provided for the password creating task; participants were instructed to create a new password in every password creation page, as follows: '*Please create a new password and confirm it. When you have finished, click on submit.*' However, when the stored data was checked, the author found that two participants created identical passwords in all five password creation tasks. Therefore, this issue was addressed in the main study by emphasising the creation of a new password that was different than the previous one, as follows: '*Create a completely different strong password and confirm it. When you have finished, click on submit.*' The second issue was raised by one participant who felt the pressure to remember each password they

created as this was repeated in every password creation page, as follows: ‘*We will ask you to recall this password in three days so it is important that you remember your new password. Try to remember it!*’ Therefore, in the main study, this issue was addressed by removing this statement from the password creation pages as it was already included in the scenario page. The data from the pilot were not included in the analysis.

8.2.5 Procedure

The same procedure as in Study 4 (see Section 6.2.5, Chapter 6) was followed in the present study. The MTurk platform was used to recruit participants and direct them to the password creation application. Each participant was assigned randomly to one of the four groups: *Group 1*, *Group 2*, *Group 3*, and *Group 4*. Three days after the password creation, participants were invited to return and recall their passwords through MTurk.

8.2.6 Data Analysis

To test for normality on all dependent measures, Kolmogorov-Smirnov and Shapiro-Wilk tests were used. The majority of dependent measures were significantly non-normal ($p < 0.05$) for both tests (except the password strength score). Therefore, non-parametric statistics were used throughout the analysis. During the data preparation process, outliers were identified and adjusted following the same method as in Study 4 (see Section 6.2.6, Chapter 6).

The data were analysed using between-participants and within-participants analyses. The between-participants analysis was used to compare participants’ performance between the four timing of presentation conditions for the password policy and creation suggestion (*before-interaction*, *during-interaction*, *after-interaction*, and *during&after-interaction*); and between the three media conditions for the password strength indicator (*graphical*, *textual*, and *graphical&textual*). Kruskal-Wallis tests (H statistic) were used for this analysis. On the other hand, the within-participants analysis was used to compare participants’ performance between the *baseline*

condition and one type of supporting feature (i.e. *baseline* and *policy*, *baseline* and *suggestion*, *baseline* and *3colour indicator*, and *baseline* and *single-colour indicator*); and between the two colour-scheme conditions (*3colour* and *single-colour*). Wilcoxon signed-ranks tests (Z statistic) were used for this analysis. Furthermore, when the dependent measures were of a frequency type, chi-square (x^2 statistics) tests were used to measure the association among the categories (i.e. number of character classes, policy compliance, suggestion compliance, password guessability, and accuracy).

A final remark regarding the data analysis is that for the password strength indicator, the interactions between the media and colour-scheme independent variables were not tested because there was no non-parametric equivalent to a two-way mixed analysis of variance.

8.3 Results

The results of this study are divided into two parts: first, the password creation results from Part I, and second, the password recall results from Part II. For Part I, 257 participants were included in the analysis: 68 in *Group 1*, 63 in *Group 2*, 61 in *Group 3*, and 65 in *Group 4*. For Part II, only 168 participants (65.37%) returned to the recall task: 47 in *Group 1*, 43 in *Group 2*, 37 in *Group 3*, and 41 in *Group 4*.

The two parts' results⁹ are presented according to the types of supporting feature, as follows: password policy in Section 8.3.1, password creation suggestion in Section 8.3.2, and password strength indicator in Section 8.3.3. Finally, the users' common practices in password creation and recall are reported in Section 8.3.4.

⁹ If half of the items in a dependent measure (e.g. user satisfaction) showed a significant effect that was not considered enough evidence to conclude that there was an effect for that dependent measure.

8.3.1 Password Policy

8.3.1.1 Password Creation

The following section presents the results regarding the usability of PCS and the strength of the created passwords when users create passwords.

8.3.1.1.1 Usability of PCS

8.3.1.1.1.1 Efficiency Measures

Table 8.4 summarises the between-participants results for the two efficiency measures, and Table 8.5 illustrates the pairwise comparison results between the four timing of presentation conditions for these measures.

Table 8.4 Mean (median) creation time and keystrokes measures between the different timing of presentation conditions for policy

	Timings of presentation				<i>P</i> value
	<i>policy-before- interaction</i> (Group 1)	<i>policy- during- interaction</i> (Group 2)	<i>policy-after- interaction</i> (Group 3)	<i>policy-during&after- interaction</i> (Group 4)	
Creation time	112.07(35.00)	84.32 (41.00)	125.92 (61.00)	102.89 (50.50)	.035
Keystrokes	23.34 (17.00)	23.13 (17.00)	39.18 (34.50)	27.23 (23.50)	.000

Creation time. There was a significant difference in the creation time between the different timing of presentation conditions for the password policy ($H(3) = 8.60$, $p = .035$). Participants in the *policy-before-interaction* and *policy-during-interaction* conditions spent significantly less time creating passwords than those in the *policy-after-interaction* condition. The pairwise comparison confirmed this difference and showed no significant difference in the creation time between the *policy-before-interaction* and *policy-during-interaction* conditions (see Table 8.5). In addition, there was a significant difference in the creation time between the *policy* and *baseline* conditions ($Z = -10.78$, $p < .001$): participants spent a significantly longer time creating passwords in the *policy* ($M = 106.23$, $Mdn = 44.00$) than in the *baseline* condition ($M = 21.24$, $Mdn = 19.00$).

Table 8.5 Pairwise comparisons of creation time and keystrokes measures between different timing of presentation conditions for policy

		<i>policy- before- interaction</i>	<i>policy- during- interaction</i>	<i>policy- after- interaction</i>	<i>policy- during&afte r-interaction</i>
Creation time	<i>policy-before-interaction</i>	-	-0.04	-32.27*	-19.57
	<i>policy-during-interaction</i>		-	-32.23*	-19.53
	<i>policy-after-interaction</i>			-	12.69
	<i>policy-during&after- interaction</i>				-
Keystrokes	<i>policy-before-interaction</i>	-	5.53	-45.64*	-9.85
	<i>policy-during-interaction</i>		-	-51.17*	-15.38
	<i>policy-after-interaction</i>			-	35.79*
	<i>policy-during&after- interaction</i>				-

Note. * denotes a significant result in pairwise comparison, $p < .05$.

Keystrokes. There was a significant difference in the number of keystrokes used to create a password between the different timing of presentation conditions for password policy ($H(3) = 18.31$, $p < .001$). Participants in the *policy-after-interaction* condition performed significantly more keystrokes than those in the other three conditions. The pairwise comparison showed a significant difference between the *policy-after-interaction* condition and the other three conditions: *policy-before-interaction*, *policy-during-interaction*, and *policy-during&after-interaction* (see Table 8.5). There was also a significant difference in the number of keystrokes between the *policy* and *baseline* conditions ($Z = -7.92$, $p < .001$): participants used significantly more keystrokes in the *policy* ($M = 28.02$, $Mdn = 23.00$) than in the *baseline* condition ($M = 16.24$, $Mdn = 14.00$).

8.3.1.1.1.2 User Satisfaction Measures

Table 8.6 summarises the between-participants results for the six user satisfaction measures.

Ease of use. There was no significant difference in the ratings of ease of use between the four timing of presentation conditions for password policy ($H(3) = 2.20$, $p = .533$). However, there was significant difference in these ratings ($Z = -4.47$, $p < .001$) between the *policy* and *baseline* conditions: presenting a password *policy* ($M = 2.47$,

$Mdn = 2.00$) was perceived to make the password creation easier compared to the *baseline* ($M = 2.04$, $Mdn = 1.00$).

Table 8.6 Mean (median) ratings of ease of use, annoyingness, helpfulness, clarity, amount of detail, and participants' confidence measures between the different timings of presentation for policy

	Timings of presentation				<i>p</i> value
	<i>policy- before- interaction</i> (Group 1)	<i>policy- during- interaction</i> (Group 2)	<i>policy- after- interaction</i> (Group 3)	<i>policy- during&after- interaction</i> (Group 4)	
Ease of use	2.50 (2.00)	2.30 (2.00)	2.66 (3.00)	2.45 (2.00)	n.s.
Annoyingness	2.41 (2.00)	2.33 (2.00)	2.59 (2.00)	2.54 (2.00)	n.s.
Helpfulness	3.16 (3.00)	3.32 (3.00)	3.18 (3.00)	3.42 (3.00)	n.s.
Clarity	3.84 (4.00)	3.98 (4.00)	3.61 (4.00)	4.03 (4.00)	n.s.
Amount of detail	2.06 (2.00)	1.97 (2.00)	2.13 (2.00)	2.29 (2.00)	n.s.
Confidence	3.53 (4.00)	3.65 (4.00)	3.49 (3.00)	3.85 (4.00)	n.s.

Annoyingness. There was no significant difference in the ratings of annoyingness between the four timing of presentation conditions for password policy ($H(3) = 1.93$, $p = .587$), or between the *policy* and *baseline* conditions ($Z = -1.87$, $p = .062$).

Helpfulness. There was no significant difference in the ratings of helpfulness between the four timing of presentation conditions for password policy ($H(3) = 2.94$, $p = .400$). However, there was significant difference in these ratings between the *policy* and *baseline* conditions ($Z = -7.67$, $p < .001$): providing a *policy* statement ($M = 3.27$, $Mdn = 3.00$) when creating a password was perceived to be more helpful than not providing one ($M = 2.65$, $Mdn = 3.00$).

Clarity. There was no significant difference in the ratings of clarity between the four timing of presentation conditions for password policy ($H(3) = 7.33$, $p = .097$). However, there was a significant difference in these ratings between the *policy* and *baseline* conditions ($Z = -1.99$, $p = .047$): participants perceived that being provided with the *policy* ($M = 3.87$, $Mdn = 4.00$) when creating a password made the password creation clearer than in the *baseline* condition ($M = 3.68$, $Mdn = 4.00$).

Amount of detail. There was no significant difference in the ratings of amount of detail between the four timing of presentation conditions for password policy ($H(3) =$

6.36, $p = .095$). However, there was a significant difference in the ratings of amount of detail between the *policy* and *baseline* conditions ($Z = -6.98$, $p < .001$): these ratings were higher in the *policy* ($M = 2.11$, $Mdn = 2.00$) than in the *baseline* condition ($M = 1.68$, $Mdn = 2.00$).

Confidence. There was no significant difference in the ratings of participants' confidence in creating passwords between the four timing of presentation conditions for password policy ($H(3) = 4.98$, $p = .174$). However, there was a significant difference in these ratings between the *policy* and *baseline* conditions ($Z = -4.36$, $p < .001$): participants felt more confident creating passwords in the *policy* ($M = 3.63$, $Mdn = 4.00$) than in the *baseline* condition ($M = 3.23$, $Mdn = 3.00$).

8.3.1.1.2 Strength of Password

8.3.1.1.2.1 Password Characteristics

Table 8.7 summarises the between-participants results for the five password characteristics measures, and Table 8.8 illustrates the pairwise comparison results between the four timing of presentation conditions for these measures.

Password length. There was a significant difference in the length of the passwords between the four timing of presentation conditions for password policy ($H(3) = 9.35$, $p = .025$): passwords created under the *policy-after-interaction* condition were longer than those created in the other three conditions. Furthermore, the pairwise comparison showed a significant difference in length between the passwords created in the *policy-after-interaction* condition and both the *policy-before-interaction* and *policy-during-interaction* conditions (see Table 8.8). However, there was no significant difference in the password length between the *policy-after-interaction* and *policy-during&after-interaction* conditions. In addition, there was a significant difference in the password length between the *policy* and *baseline* conditions ($Z = -9.63$, $p < .001$): participants created longer passwords in the *policy* ($M = 13.08$, $Mdn = 13.00$) than in the *baseline* condition ($M = 10.76$, $Mdn = 10.00$).

Table 8.7 Mean (median) Password length, number of digits, number of uppercase letters, number of lowercase letters and number of symbols measures between different timing of presentation for policy

	Timings of presentation				<i>p</i> value
	<i>policy-before-interaction</i> (Group 1)	<i>policy-during-interaction</i> (Group 2)	<i>policy-after-interaction</i> (Group 3)	<i>policy-during&after-interaction</i> (Group 4)	
Password length	12.79 (12.00)	12.39 (12.00)	13.92 (14.00)	13.25 (13.00)	.025
Number of digits	2.84 (3.00)	3.11 (3.00)	3.36 (3.00)	3.46 (4.00)	n.s.
Number of uppercase	1.66 (1.00)	1.20 (1.00)	1.10 (1.00)	1.25 (1.00)	n.s.
Number of lowercase	6.64 (7.00)	6.17 (6.00)	7.47 (7.00)	7.12 (7.00)	n.s.
Number of symbols	0.82 (1.00)	1.01 (1.00)	0.96 (1.00)	0.72 (1.00)	n.s.

Table 8.8 Pairwise comparisons of password length measures between timing of presentation conditions for policy

		<i>policy-before-interaction</i>	<i>policy-during-interaction</i>	<i>policy-after-interaction</i>	<i>policy-during&after-interaction</i>
Password length	<i>policy-before-interaction</i>	-	12.43	-26.31*	-11.37
	<i>policy-during-interaction</i>		-	-38.75*	-23.81
	<i>policy-after-interaction</i>			-	14.94
	<i>policy-during&after-interaction</i>				-

Note. * denotes a significant result in pairwise comparison, $p < .05$.

Number of digits. There was no significant difference in the number of digits used in passwords between the four timing of presentation conditions for password policy ($H(3) = 4.54$, $p = .209$). However, there was a significant difference in this measure between the *policy* and *baseline* conditions ($Z = -3.89$, $p < .001$): passwords created in the *policy* condition ($M = 3.19$, $Mdn = 3.00$) had more digits than in the *baseline* condition ($M = 2.68$, $Mdn = 3.00$).

Number of uppercase letters. There was no significant difference in the number of uppercase letters used in passwords between the four timing of presentation conditions for password policy ($H(3) = 5.40$, $p = .144$). However, there was a significant difference in this measure between the *policy* and *baseline* conditions ($Z = -5.21$, $p < .001$): passwords created in the *policy* condition ($M = 1.31$, $Mdn = 1.00$) had more uppercase letters than in the *baseline* condition ($M = 0.92$, $Mdn = 1.00$).

Number of lowercase letters. There was no significant difference in the number of lowercase letters used in passwords between the four timing of presentation conditions for password policy ($H(3) = 4.47, p = .215$). However, there was a significant difference in this measure between the *policy* and *baseline* conditions ($Z = -2.48, p = .013$): passwords created in the *policy* condition ($M = 6.84, Mdn = 7.00$) had more lowercase letters than in the *baseline* condition ($M = 6.16, Mdn = 6.00$).

Number of symbols. There was no significant difference in the number of symbols used in passwords between the four timing of presentation conditions for password policy ($H(3) = 2.19, p = .534$). However, there was a significant difference in this measure between the *policy* and *baseline* conditions ($Z = -8.48, p < .001$): passwords created in the *policy* condition ($M = 0.87, Mdn = 1.00$) had more symbols than in the *baseline* condition ($M = 0.38, Mdn = 0.00$).

Number of password character classes. With regard to the number of different character classes in passwords, there were significant differences in all four timing of presentation conditions: *policy-before-interaction* (Group 1: $\chi^2(3) = 46.47, p < .001$), *policy-during-interaction* (Group 2: $\chi^2(3) = 36.87, p < .001$), *policy-after-interaction* (Group 3: $\chi^2(2) = 18.33, p < .001$), and *policy-during&after-interaction* (Group 4: $\chi^2(3) = 39.19, p < .001$). The distribution of the password character classes across the different timings is shown in Figure 8.12. The *policy-after-interaction* condition had the highest percentage of passwords (57.38%) using all four character classes, followed by *policy-during&after-interaction* (53.85%), *policy-before-interaction* (52.94%), and *policy-during-interaction* (50.79%). Furthermore, passwords created in the *policy* condition had significantly more character classes than those in the *baseline* condition ($Z = -9.94, p < .001$). In the *policy* condition, the majority of passwords included four character classes (53.70%), whereas in the *baseline* condition, most passwords included three (35.41%) or two (30.74%) character classes.

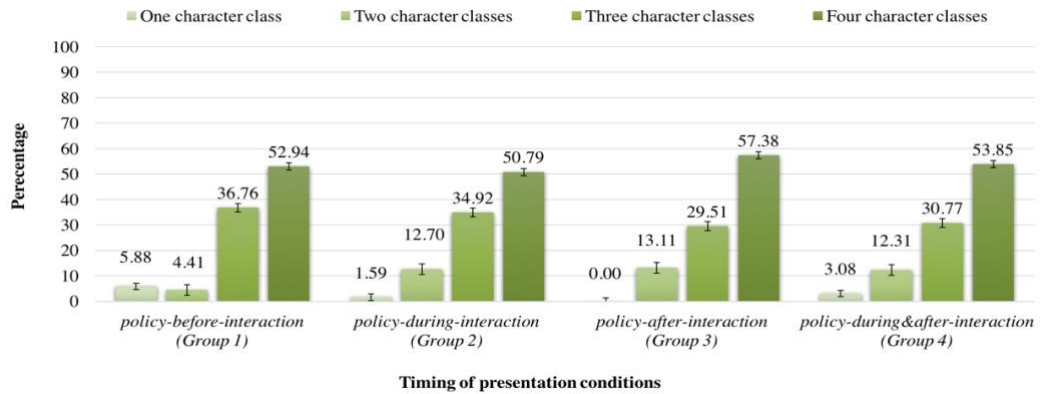


Figure 8.12 Percentage of password character classes across the four timings of presentation for policy

Policy compliance. There were significant differences in the number of passwords that followed the given policy in three timing of presentation conditions: *policy-before-interaction* (Group 1: $\chi^2(1) = 11.53$, $p = .001$), *policy-after-interaction* (Group 3: $\chi^2(1) = 10.25$, $p = .001$), and *policy-during&after-interaction* (Group 4: $\chi^2(1) = 14.79$, $p < .001$), but not *policy-during-interaction* (Group 2: $\chi^2(1) = 1.92$, $p = .166$). The distribution of the policy compliance across the different timing of presentation conditions is shown in Figure 8.13. The *policy-during&after-interaction* condition had the highest percentage of passwords (73.85%) that followed the policy, followed by *policy-before-interaction* (70.59%) and *policy-after-interaction* (70.49%).

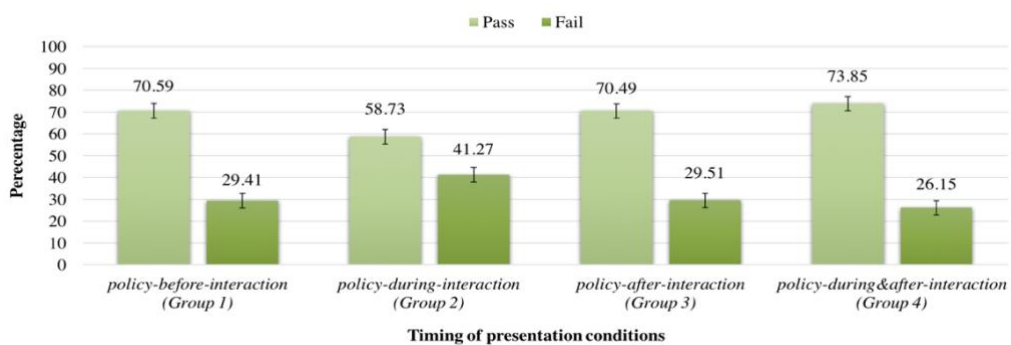


Figure 8.13 Percentage of password compliance across the four timings of presentation for policy

8.3.1.1.2.2 Password Guessability

Out of the 257 passwords, 176 complied with the password policy chosen for this study (at least 12 characters and at least three character classes). The ability to guess the 176 passwords was examined using five cracking approaches (based on Ur et al. (2015) as another measure for password strength. Overall, just under three-quarters of the passwords (129, 73.30%) were not guessed at all, whereas 47 (26.70%) were guessed by at least one of the five approaches. The distribution of the password guessability across the different timing of presentation conditions is shown in Figure 8.14.

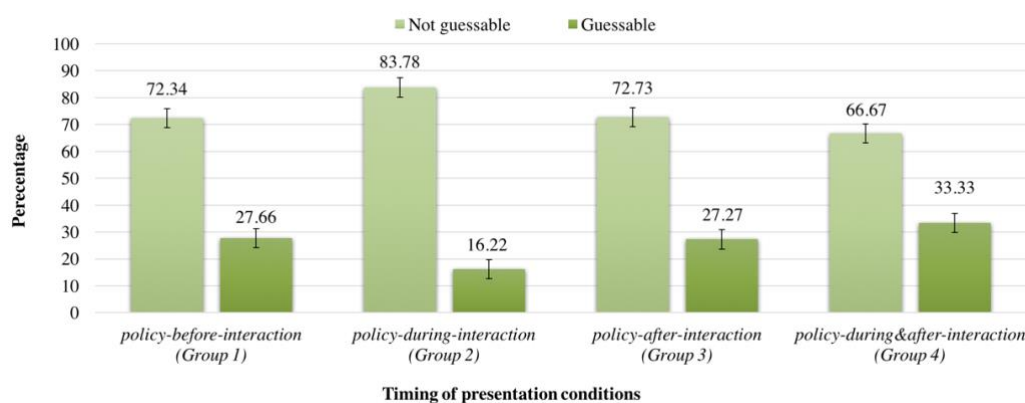


Figure 8.14 Percentage of password guessability across the four timings of presentation for policy. With regard to the four timing of presentation conditions, there were significant differences between the number of passwords that were guessable and those that were non-guessable in all four timing of presentation conditions: *policy-before-interaction* (Group 1: $\chi^2(1) = 9.38$, $p = .002$), *policy-during-interaction* (Group 2: $\chi^2(1) = 16.89$, $p < .001$), *policy-after-interaction* (Group 3: $\chi^2(1) = 9.09$, $p = .003$) and *policy-during&after-interaction* (Group 4: $\chi^2(1) = 5.33$, $p = .021$). The *policy-during-interaction* condition had the highest percentage of passwords (83.78%) that were not guessable, whereas the *policy-during&after-interaction* (66.67%) had the lowest percentage.

8.3.1.2 Password Recall

The following section presents the results regarding the usability of PCS when users recall passwords.

8.3.1.2.1 Efficiency Measures

Recall time. There was no significant difference in the recall time between the four timing of presentation conditions ($H(3) = 3.39, p = .335$) for password policy. Figure 8.15 shows the mean recall times across the four timings.

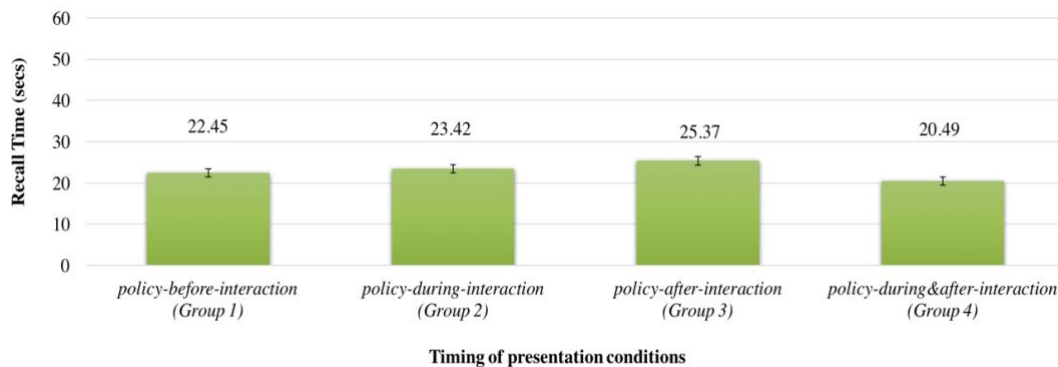


Figure 8.15 Mean recall times across the four timing of presentation conditions for policy

However, there was a significant difference in the recall time between the *policy* and *baseline* conditions ($Z = -9.56, p < .001$). The recall time for passwords created in the *policy* condition ($M = 22.86, Mdn = 21.00$) was shorter than for those created in the *baseline* condition ($M = 39.80, Mdn = 38.00$).

8.3.1.2.2 Effectiveness Measures

Accuracy. There was a significant difference in the accuracy of recalling passwords in all four timing of presentation conditions: *policy-before-interaction* (Group 1: $\chi^2(1) = 4.79, p = .029$), *policy-during-interaction* (Group 2: $\chi^2(1) = 6.72, p = .010$), *policy-after-interaction* (Group 3: $\chi^2(1) = 4.57, p = .033$), and *policy-during&after-interaction* (Group 4: $\chi^2(1) = 5.49, p = .019$). The percentages of accuracy in password recall across the four timing of presentation conditions are shown in Figure 8.16.

The highest percentage of successful recall (34.04%) was in the *policy-before-interaction* condition, and the highest rate of unsuccessful recall (69.77%) was in the *policy-during-interaction* condition. In addition, there was a significant difference in the accuracy of password recall between the *policy* and *baseline* conditions ($Z = -3.12, p = .002$): it was lower in the *policy* (32.14%) than in the *baseline* (43.45%) condition.

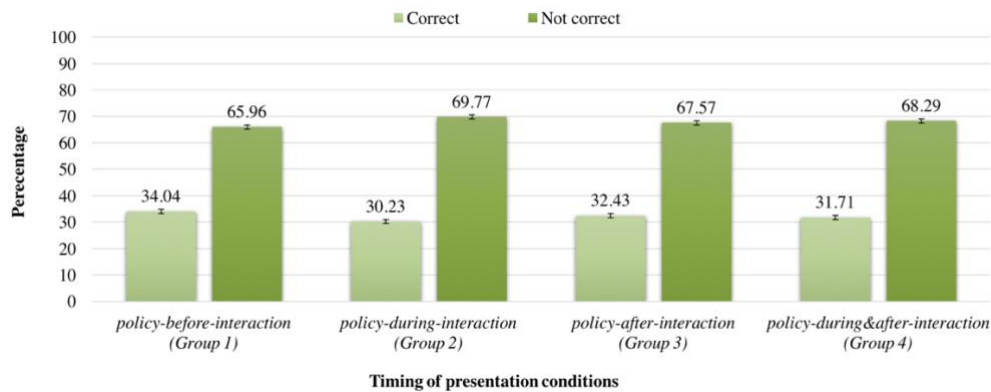


Figure 8.16 Percentages of accuracy in recalling passwords across the four timing of presentation conditions for policy

8.3.1.2.3 User Satisfaction Measures

Confidence. There was no significant difference in the ratings of participants' confidence in recalling the correct passwords between the different timing of presentation conditions for password policy ($H(3) = 6.20, p = .102$). The mean ratings of this measure across the four timings are shown in Figure 8.17.

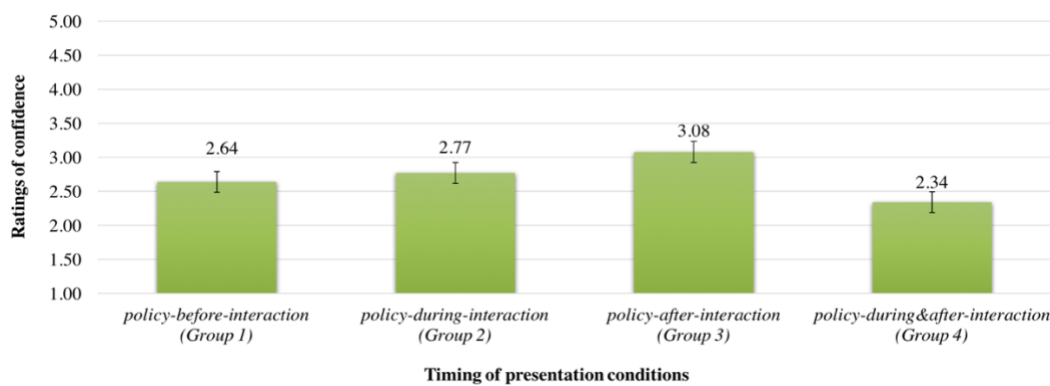


Figure 8.17 Mean ratings of participants' confidence across the four timing of presentation conditions for policy

However, there was a significant difference in the ratings of participants' confidence between the *policy* and *baseline* conditions ($Z = -3.33, p = .001$): participants felt significantly less confident in recalling the correct passwords in the *policy* ($M = 2.70, Mdn = 3.00$) than in the *baseline* condition ($M = 3.01, Mdn = 3.00$).

8.3.1.3 Summary

For the password creation process, the four timings of presentation of password policy did differ significantly in terms of password creation time, number of keystrokes, password length, number of password character classes, policy compliance, and password guessability. Participants used significantly less time and effort to create passwords in the *policy-before-interaction* and *policy-during-interaction* conditions. However, participants in the *policy-before-interaction* condition created significantly more compliant passwords than those in the *policy-during-interaction* condition. Furthermore, the passwords created in the *policy-after-interaction* and *policy-during&after-interaction* conditions were significantly longer, with all four character classes being used, than those created in other conditions. The majority of passwords in all four timings of presentation were not guessable, with participants in the *policy-during-interaction* creating the highest percentage of non-guessable passwords.

The PCS usability and password strength between the *policy* and *baseline* conditions did differ significantly in terms of the efficiency of the PCSs, level of user satisfaction (except annoyingness) with the PCSs, and password characteristics. Providing a policy statement made the password creation process less efficient, but it did improve the level of user satisfaction and the password characteristic.

For the password recall, the four timing of presentation conditions did not differ significantly in terms of recall time and participants' confidence. However, the level of accuracy did differ significantly in all four timings: successful recall rates were low compared to the unsuccessful ones.

Finally, a significant difference was found between the *policy* and *baseline* conditions in terms of the recall time, accuracy, and participants' confidence when recalling passwords. Participants spent significantly less time recalling passwords in the *policy* conditions, but they did not recall passwords correctly and they also felt less confident than in the *baseline* condition.

8.3.2 Password Creation Suggestion

8.3.2.1 Password Creation

The following section presents the results regarding the usability of PCS and the strength of the created passwords when users create passwords.

8.3.2.1.1 Usability of PCS

8.3.2.1.2 Efficiency Measures

Table 8.9 summarises the between-participants results for the two efficiency measures, and Table 8.10 illustrates the pairwise comparison results between the four timing of presentation conditions for these measures.

Creation time. There was a significant difference in the creation time ($H(3) = 10.79$, $p = .013$) between the different timing of presentation conditions for password creation suggestion: specifically, participants in the *suggestion-after-interaction* condition spent significantly more time creating passwords than those in the other three conditions. However, the pairwise comparison showed no significant difference in this measure between the *suggestion-during&after-interaction* and both the *suggestion-before-interaction* and *suggestion-during-interaction* conditions (see Table 8.10). In addition, there was a significant difference in the measure between the *suggestion* and *baseline* conditions ($Z = -3.01$, $p = .003$): participants spent a significantly longer time creating passwords in the *suggestion* ($M = 89.04$, $Mdn = 44.00$) than in the *baseline* condition ($M = 21.24$, $Mdn = 19.00$).

Table 8.9 Mean (median) creation time and keystrokes measures between the different timing of presentation conditions for suggestion

	Timings of presentation				<i>p</i> value
	<i>suggestion-before-interaction</i> (Group 1)	<i>suggestion-during-interaction</i> (Group 2)	<i>suggestion-after-interaction</i> (Group 3)	<i>suggestion-during&after-interaction</i> (Group 4)	
Creation time	92.99 (22.00)	85.28 (28.00)	108.57 (39.00)	70.23 (26.00)	.013
Keystrokes	14.73 (12.00)	20.54 (17.50)	25.53 (24.50)	15.67 (14.00)	.000

Table 8.10 Pairwise comparisons of creation time and keystrokes measures between different timing of presentation conditions for suggestion

		<i>suggestion- before- interaction</i>	<i>suggestion- during- interaction</i>	<i>suggestion- after- interaction</i>	<i>suggestion- during&afte r-interaction</i>
Creation time	<i>suggestion-before-interaction</i>	-	-16.37	-42.43*	-14.00
	<i>suggestion-during-interaction</i>		-	-26.06	-2.37
	<i>suggestion-after-interaction</i>			-	28.43*
	<i>suggestion-during&after- interaction</i>				-
Keystrokes	<i>suggestion-before-interaction</i>	-	-40.56*	-64.65*	-10.67
	<i>suggestion-during-interaction</i>		-	-24.08	29.89*
	<i>suggestion-after-interaction</i>			-	53.97*
	<i>suggestion-during&after- interaction</i>				-

Note. * denotes a significant result in pairwise comparison, $p < .05$.

Keystrokes. There was a significant difference in the number of keystrokes used to create a password between the different timing of presentation conditions for password creation suggestion ($H(3) = 30.13$, $p < .001$): participants in the *suggestion-after-interaction* and *suggestion-during-interaction* conditions performed significantly more keystrokes than in the other two conditions. The pairwise comparison confirmed this difference, but showed no significant difference between the *suggestion-before-interaction* and *suggestion-during&after-interaction* conditions (see Table 8.10). In addition, there was a significant difference in this measure between the *suggestion* and *baseline* conditions ($Z = -3.01$, $p = .003$): participants used significantly more keystrokes in the *suggestion* ($M = 18.94$, $Mdn = 16.00$) than in the *baseline* condition ($M = 16.24$, $Mdn = 14.00$).

8.3.2.1.3 User Satisfaction Measures

Table 8.11 summarises the between-participants results for the six user satisfaction measures, and Table 8.12 illustrates the pairwise comparison results between the four timing of presentation conditions for these measures.

Ease of use. There was no significant difference in the ratings of ease of use between the four timing of presentation conditions for password creation suggestion ($H(3) = 5.46$, $p = .141$). However, there were significant differences in these ratings between

the *suggestion* and *baseline* conditions ($Z = -4.04, p < .001$). Participants perceived that providing the *suggestion* ($M = 2.33, Mdn = 2.00$) made the password creation easier than in the *baseline* condition ($M = 2.04, Mdn = 1.00$).

Table 8.11 Mean (median) ratings of ease of use, annoyingness, helpfulness, clarity, amount of detail, and participants’ confidence measures between the different timings of presentation for suggestion

	Timings of presentation				<i>p</i> value
	<i>suggestion-before-interaction</i> (Group 1)	<i>suggestion-during-interaction</i> (Group 2)	<i>suggestion-after-interaction</i> (Group 3)	<i>suggestion-during&after-interaction</i> (Group 4)	
Ease of use	2.31 (2.00)	2.10 (1.00)	2.49 (2.00)	2.42 (2.00)	n.s.
Annoyingness	2.71 (2.00)	2.41 (2.00)	2.56 (3.00)	2.55 (2.00)	n.s.
Helpfulness	2.99 (3.00)	3.05 (3.00)	2.93 (3.00)	2.95 (3.00)	n.s.
Clarity	3.76 (4.00)	3.90 (4.00)	3.57 (3.00)	3.91 (4.00)	n.s.
Amount of detail	2.01 (2.00)	1.68 (2.00)	2.07 (2.00)	2.00 (2.00)	.027
Confidence	3.29 (3.00)	3.83 (4.00)	3.43 (3.00)	3.65 (4.00)	.016

Annoyingness. There was no significant difference in the ratings of annoyingness between the four timing of presentation conditions for password creation suggestion ($H(3) = 1.54, p = .674$), or between the *suggestion* and *baseline* conditions ($Z = -0.78, p = .435$).

Table 8.12 Pairwise comparisons of amount of detail and confidence measures between different timing of presentation conditions for suggestion

		<i>suggestion-before-interaction</i>	<i>suggestion-during-interaction</i>	<i>suggestion-after-interaction</i>	<i>suggestion-during&after-interaction</i>
Amount of detail	<i>suggestion-before-interaction</i>	-	29.54*	-3.94	1.34
	<i>suggestion-during-interaction</i>		-	-33.48*	-28.21*
	<i>suggestion-after-interaction</i>			-	5.27
	<i>suggestion-during&after-interaction</i>				-
Confidence	<i>suggestion-before-interaction</i>	-	-37.19*	-8.36	-23.96
	<i>suggestion-during-interaction</i>		-	28.84*	13.23
	<i>suggestion-after-interaction</i>			-	-15.60
	<i>suggestion-during&after-interaction</i>				-

Note. * denotes a significant result in pairwise comparison, $p < .05$.

Helpfulness. There was no significant difference in the ratings of helpfulness between the four timing of presentation conditions for password creation suggestion ($H(3) =$

1.34, $p = .720$). However, there was a significant difference in the ratings of helpfulness between the *suggestion* and *baseline* conditions ($Z = -4.92$, $p < .001$): providing a *suggestion* statement ($M = 2.98$, $Mdn = 3.00$) when creating password was perceived to be more helpful than not providing one ($M = 2.65$, $Mdn = 3.00$).

Clarity. There was no significant difference in the ratings of clarity between the four timing of presentation conditions for password creation suggestion ($H(3) = 3.67$, $p = .299$), or between the *suggestion* and *baseline* conditions ($Z = -1.42$, $p = .157$).

Amount of detail. There was a significant difference in the ratings of amount of detail between the four timing of presentation conditions for password creation suggestion ($H(3) = 9.14$, $p = .027$). Namely, the ratings were significantly lower in the *suggestion-during-interaction* condition than in the other three conditions. The pairwise comparison confirmed this difference. It also showed no significant difference in ratings of amount of detail between the other three conditions: *suggestion-before-interaction*, *suggestion-after-interaction*, and *suggestion-during&after-interaction* (see Table 8.12). In addition, there were significant differences in these ratings between the *suggestion* and *baseline* conditions ($Z = -4.87$, $p < .001$): participants gave the measure a higher rating in the *suggestion* condition ($M = 1.94$, $Mdn = 2.00$) than in the *baseline* condition ($M = 1.68$, $Mdn = 2.00$).

Confidence. There was a significant difference in the ratings of participants' confidence in creating passwords between the four timing of presentation conditions for password creation suggestion ($H(3) = 10.29$, $p = .016$): participants in the *suggestion-during-interaction* condition felt significantly more confident than those in the *suggestion-before-interaction* and *suggestion-after-interaction* conditions when creating passwords. The pairwise comparison confirmed this difference, but showed no significant difference between the *suggestion-during-interaction* and *suggestion-during&after-interaction* conditions (see Table 8.12). In addition, there was a significant difference in this measure between the *suggestion* and *baseline* conditions ($Z = -3.81$, $p < .001$). Namely, participants felt more confident creating passwords in the *suggestion* ($M = 3.54$, $Mdn = 4.00$) than in the *baseline* condition ($M = 3.23$, $Mdn = 3.00$).

8.3.2.1.4 Strength of Password

8.3.2.1.5 Password Characteristics

Table 8.13 summarises the between-participants results for the five password characteristics measures, and Table 8.14 illustrates the pairwise comparison results between the four timing of presentation conditions for these measures.

Table 8.13 Mean (median) password length, number of digits, number of uppercase letters, number of lowercase letters, and number of symbols measures between different timings of presentation for suggestion

	Timings of presentation				<i>p</i> value
	<i>suggestion-before-interaction</i> (Group 1)	<i>suggestion-during-interaction</i> (Group 2)	<i>suggestion-after-interaction</i> (Group 3)	<i>suggestion-during&after-interaction</i> (Group 4)	
Password length	9.95 (10.00)	11.33 (11.00)	11.95 (12.00)	10.99 (11.00)	.002
Number of digits	2.10 (2.00)	2.79 (2.00)	2.84 (3.00)	2.35 (2.00)	n.s.
Number of uppercase	0.86 (1.00)	0.73 (1.00)	1.15 (1.00)	0.72 (1.00)	n.s.
Number of lowercase	5.37 (6.00)	6.20 (6.00)	6.31 (6.00)	6.43 (6.00)	n.s.
Number of symbols	1.15 (1.00)	1.07 (1.00)	1.10 (1.00)	0.95 (1.00)	n.s.

Table 8.14 Pairwise comparisons of password length measures between different timing of presentation conditions for suggestion

		<i>suggestion-before-interaction</i>	<i>suggestion-during-interaction</i>	<i>suggestion-after-interaction</i>	<i>suggestion-during&after-interaction</i>
Password length	<i>suggestion-before-interaction</i>	-	-39.91*	-46.62*	-26.30*
	<i>suggestion-during-interaction</i>		-	-6.72	13.61
	<i>suggestion-after-interaction</i>			-	20.32
	<i>suggestion-during&after-interaction</i>				-

Note. * denotes a significant result in pairwise comparison, $p < .05$.

Password length. There was a significant difference in the length of the passwords between the four timing of presentation conditions for password creation suggestion ($H(3) = 15.28$, $p = .002$). Specifically, passwords created under the *suggestion-after-interaction* condition were longer than those created in the other three conditions. Furthermore, the pairwise comparison showed a significant difference between the passwords created in the *suggestion-before-interaction* condition and all other

conditions (see Table 8.14). However, there was no significant difference in the password length between the *suggestion* and *baseline* conditions ($Z = -1.33$, $p = .185$).

Number of digits. There was no significant difference in the number of digits used in passwords between the four timing of presentation conditions for password creation suggestion ($H(3) = 6.36$, $p = .096$), or between the *suggestion* and *baseline* conditions ($Z = -1.43$, $p = .153$).

Number of uppercase letters. There was no significant difference in the number of uppercase letters used in passwords between the four timing of presentation conditions for password creation suggestion ($H(3) = 3.35$, $p = .341$), or between the *suggestion* and *baseline* conditions ($Z = -0.88$, $p = .380$).

Number of lowercase letters. There was no significant difference in the number of lowercase letters used in passwords between the four timing of presentation conditions for password creation suggestion ($H(3) = 3.27$, $p = .352$), or between the *suggestion* and *baseline* conditions ($Z = -0.81$, $p = .419$).

Number of symbols. There was no significant difference in the number of symbols used in passwords between the four timing of presentation conditions for password creation suggestion ($H(3) = 2.34$, $p = .505$). However, there was a significant difference in this measure between the *suggestion* and *baseline* conditions ($Z = -10.78$, $p < .001$): passwords created in the *suggestion* condition ($M = 1.07$, $Mdn = 1.00$) had more symbols than in the *baseline* condition ($M = 0.38$, $Mdn = 0.00$).

Number of password character classes. Regarding the number of different character classes in passwords, there were significant differences in all four timing of presentation conditions: *suggestion-before-interaction* (Group 1: $\chi^2(3) = 39.88$, $p < .001$), *suggestion-during-interaction* (Group 2: $\chi^2(3) = 23.54$, $p < .001$), *suggestion-after-interaction* (Group 3: $\chi^2(3) = 25.23$, $p < .001$), and *suggestion-during&after-interaction* (Group 4: $\chi^2(3) = 22.69$, $p < .001$). Figure 8.18 shows the distribution of the password character classes across the different timing of presentation conditions.

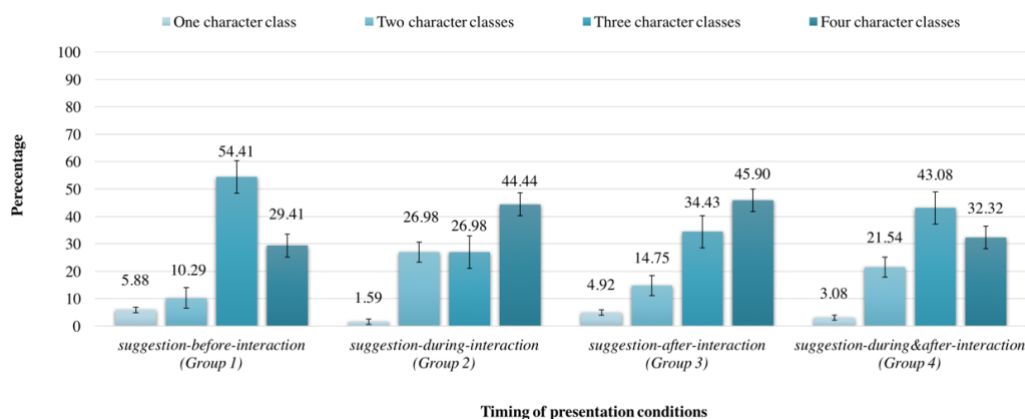


Figure 8.18 Percentage of password character classes across the four timings of presentation for suggestion

The *suggestion-after-interaction* condition had the highest percentage of passwords (45.90%) that contained four character classes, followed by *suggestion-during-interaction* (44.44%). On the other hand, the other two timings of presentation had the highest percentage of passwords containing three character classes: *suggestion-before-interaction* with 54.41% and *suggestion-during&after-interaction* with 45.90%. Furthermore, passwords created in the *suggestion* condition had significantly more character classes than in the *baseline* condition ($Z = -7.58, p < .001$). In the *suggestion* condition, most passwords included three (40.08%) or four (37.74%) character classes, whereas in the *baseline* condition, most included three (35.41%) or two (30.74%).

Suggestion compliance. There was a significant difference in the number of passwords that followed the given suggestion and those did not in two timing of presentation conditions: *suggestion-before-interaction* (Group 1: $\chi^2(1) = 5.88, p = .015$) and *suggestion-after-interaction* (Group 3: $\chi^2(1) = 8.67, p = .003$), but not in the *suggestion-during-interaction* (Group 2: $\chi^2(1) = 1.92, p = .166$) and *suggestion-during&after-interaction* (Group 4: $\chi^2(1) = 3.46, p = .063$) conditions. The distribution of the suggestion compliance across the different timing of presentation conditions is shown in Figure 8.19. The *suggestion-after-interaction* condition had the highest percentage of passwords that followed the suggestion, 68.85%, followed by *suggestion-before-interaction* with 64.71%.

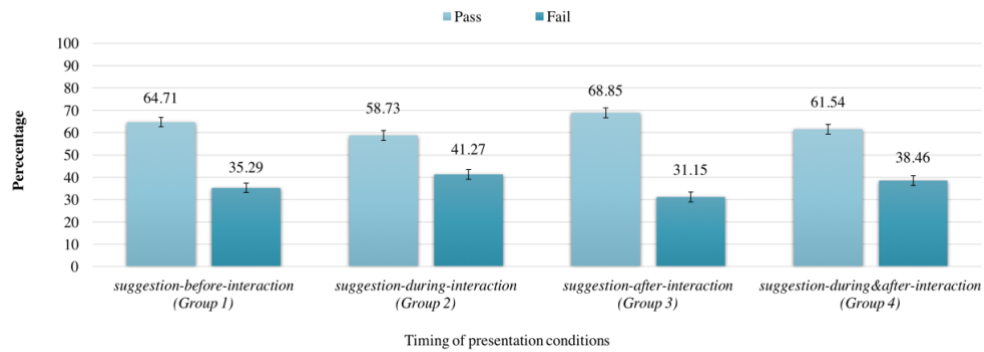


Figure 8.19 Percentage of password compliance across the four timing of presentation for suggestion **Symbols provision**. There was a significant difference in the number of passwords including the provided symbols and those did not in two timing of presentation conditions: *suggestion-during-interaction* (Group 2: $\chi^2(1) = 9.92$, $p = .002$) and *suggestion-after-interaction* (Group 3: $\chi^2(1) = 5.92$, $p = .015$), but not in the *suggestion-before-interaction* (Group 1: $\chi^2(1) = 2.88$, $p = .090$) and *suggestion-during&after-interaction* (Group 4: $\chi^2(1) = 1.86$, $p = .172$) conditions. The percentage of passwords that included at least one of the symbols in the *suggestion-during-interaction* condition was 69.84%, and 65.57% in the *suggestion-after-interaction* condition, as shown in Figure 8.20.

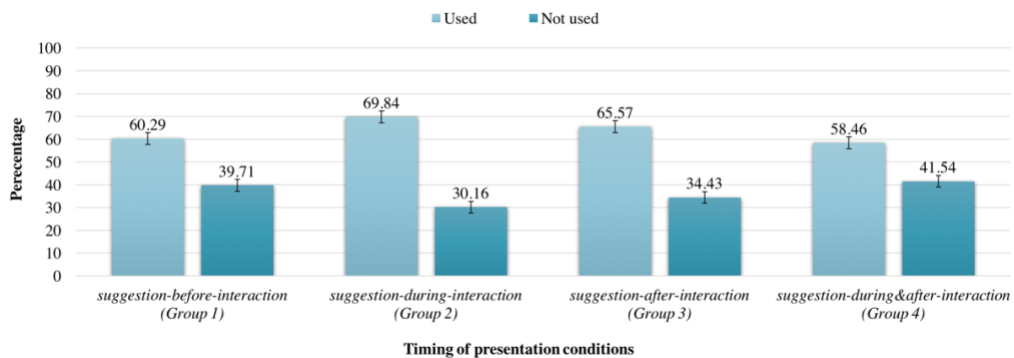


Figure 8.20 Percentage of passwords including the given symbols across the four timings of presentation

8.3.2.1.6 Password Guessability

Since participants were not forced to follow any password policy, all 257 passwords were examined to check the ability to guess them using five cracking approaches (based on Ur et al., 2015). Overall, more than half of the passwords (142, 55.25%)

were guessed by at least one approach, whereas 115 passwords (44.75%) were not guessed at all. Figure 8.21 shows the distribution of password guessability across the different timing of presentation conditions.

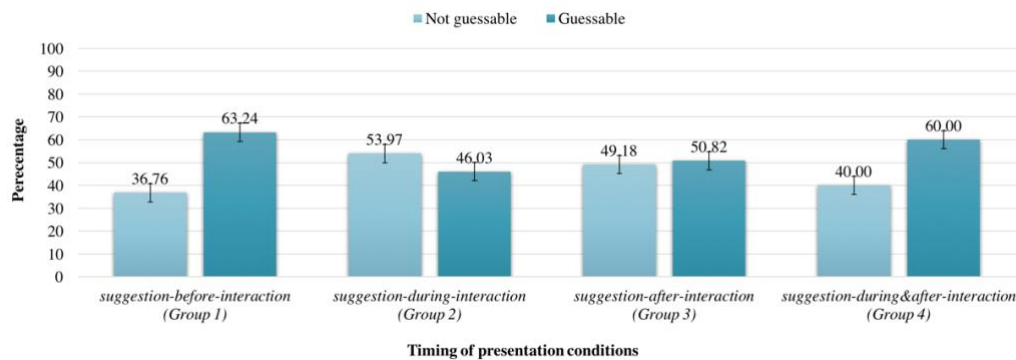


Figure 8.21 Percentage of password guessability across the four timings of presentation for suggestion. There was only a significant difference between the number of passwords that were guessable and those that were not in one timing of presentation condition: *suggestion-before-interaction* (Group 1: $\chi^2(1) = 4.77$, $p = .029$). There was no significant difference in the other three conditions: *suggestion-during-interaction* (Group 2: $\chi^2(1) = 0.40$, $p = .529$), *suggestion-after-interaction* (Group 3: $\chi^2(1) = 0.02$, $p = .898$), and *suggestion-during&after-interaction* (Group 4: $\chi^2(1) = 2.60$, $p = .107$). The highest percentage (63.24%) of guessable passwords was in the *suggestion-before-interaction* condition.

8.3.2.2 Password Recall

The following section presents the results regarding the usability of PCS when users recall passwords.

8.3.2.2.1 Efficiency Measures

Recall time. There was no significant difference in the recall time between the four timing of presentation conditions ($H(3) = 1.59$, $p = .771$) for password creation suggestion. Figure 8.22 shows the mean recall times across the four conditions. However, there was a significant difference in the recall time between the *suggestion* and *baseline* conditions ($Z = -10.04$, $p < .001$). The recall time in the *suggestion*

condition ($M = 20.80$, $Mdn = 18.00$) was shorter than in the *baseline* condition ($M = 39.80$, $Mdn = 38.00$).

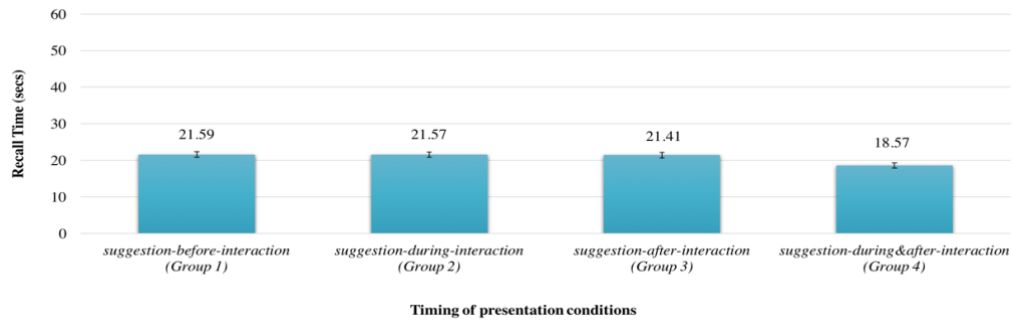


Figure 8.22 Mean recall times across the four timing of presentation conditions for suggestion

8.3.2.2.2 Effectiveness Measures

Accuracy. There was a significant difference in the accuracy of recalling passwords in two timing of presentation conditions: *suggestion-before-interaction* (Group 1: $\chi^2(1) = 9.38$, $p = .002$) and *suggestion-during&after-interaction* (Group 4: $\chi^2(1) = 7.05$, $p = .008$), but not in the *suggestion-during-interaction* (Group 2: $\chi^2(1) = 2.81$, $p = .093$) and *suggestion-after-interaction* (Group 3: $\chi^2(1) = 2.19$, $p = .139$). Figure 8.23 shows the percentage of accuracy in password recall across the conditions.

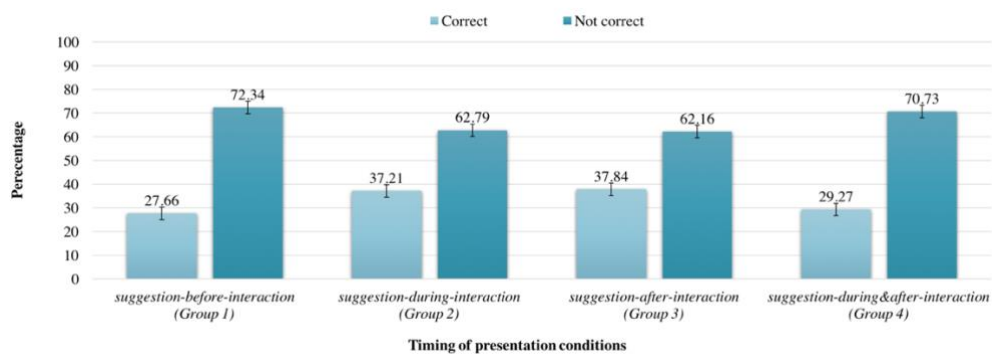


Figure 8.23 Percentages of accuracy in recalling passwords across timing of presentation conditions for suggestion

The successful recall rate was higher in the *suggestion-during&after-interaction* (29.27%) than in the *suggestion-before-interaction* condition (27.66%). In addition, there was a significant difference in the accuracy of recalling passwords between the *suggestion* and *baseline* conditions ($Z = -3.09$, $p = .002$): the accuracy of recalling

passwords correctly was lower in the *suggestion* (32.74%) than in the *baseline* (43.45%) condition.

8.3.2.2.3 User Satisfaction Measures

Confidence. There was no significant difference in the ratings of participants' confidence in recalling the correct passwords between different timing of presentation conditions for the password creation suggestion ($H(3) = 3.17, p = .367$). Figure 8.24 shows the mean ratings of this measure across the conditions. However, there was a significant difference in the measure between the *suggestion* and *baseline* conditions ($Z = -3.64, p < .001$): participants felt significantly less confident in recalling the correct passwords in the *suggestion* ($M = 2.70, Mdn = 3.00$) than in the *baseline* ($M = 3.01, Mdn = 3.00$) condition.

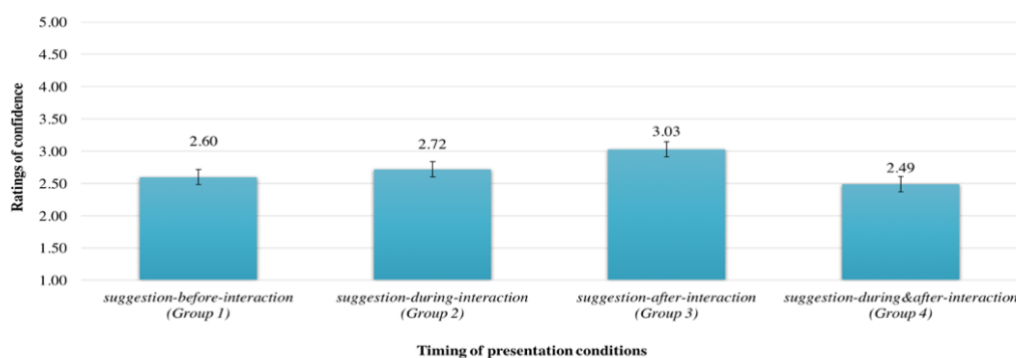


Figure 8.24 Mean ratings of participants' confidence between four timing of presentation conditions for suggestion

8.3.2.3 Summary

For the password creation, the four timings of presentation of password creation suggestion differed significantly in the creation time, number of keystrokes, PCSs' amount of detail, participants' confidence in creating passwords, password length, number of password character classes, suggestion compliance, and symbols provision. Participants used significantly less time and effort to create passwords in all of the timing of presentation conditions than in the *suggestion-after-interaction* condition. The ratings of amount of detail were low in the *suggestion-during-interaction* condition, yet participants felt more confident creating a password in this condition and in the *suggestion-during&after-interaction* condition than in the other conditions.

The passwords created in all of the timing of presentation conditions (except *suggestion-before-interaction*) were long and used four character classes. However, the *suggestion-during-interaction* and *suggestion-after-interaction* conditions had the highest percentage of passwords that included all four character classes; they also had the highest percentage of passwords that included at least one of the shown symbols. Most compliant passwords were created in the *suggestion-after-interaction* and *suggestion-before-interaction* conditions. Overall, more than half of the passwords were guessed by at least one of the five approaches; in particular, the *suggestion-before-interaction* condition had the highest percentage of guessable passwords.

The usability of PCSs and passwords strength did differ significantly between the *suggestion* and *baseline* conditions in the efficiency of the PCSs, level of user satisfaction (except annoyingness and clarity) with the PCSs, and password characteristics (only number of symbols and number of character classes). Providing a suggestion statement made the password creation process less efficient, but it did improve the level of user satisfaction and the password characteristics.

For the password recall, the four timing of presentation conditions did not differ significantly in the recall time and participants' confidence. However, the level of accuracy differed significantly in the *suggestion-before-interaction* and *suggestion-during&after-interaction*: successful recall rates were low compared to unsuccessful ones.

Finally, a significant difference was found between the *suggestion* and *baseline* conditions in terms of the recall time, accuracy, and participants' confidence when recalling passwords. Participants spent significantly less time recalling passwords in the *suggestion* conditions, but they did not recall correct passwords and felt less confident.

8.3.3 Password Strength indicator

For the password strength indicator, as discussed in Section 8.2.1, the participants in Group 4 (65 participants) were assigned to one of the three media between-participant conditions: *graphical*, *textual*, and *graphical&textual*. To this end, 25 participants

were placed in *Group 1*, 20 in *Group 2*, and finally 20 in *Group 3*. Therefore, three groups of participants were included in the analysis for Part I: 93 participants in *Group 1*, 83 in *Group 2*, 81 in *Group 3*. The distribution of the 168 participants who returned to the recall for Part II, was as follows: 65 participants in *Group 1*, 52 in *Group 2*, and 51 in *Group 3*.

8.3.3.1 Password Creation

The following section presents the results regarding the usability of PCS and the strength of the created passwords when users create passwords.

8.3.3.1.1 Usability of PCS

8.3.3.1.1.1 Efficiency Measures

Table 8.15 summarises the between-participants (in both colour-scheme conditions) results for the two efficiency measures, and Table 8.16 illustrates the pairwise comparison results between the three media conditions for these measures.

Table 8.15 Mean (median) creation time and keystrokes measures for the six indicator conditions

		Media					
		graphical (Group 1)	textual (Group 2)	graphical&tex tual (Group 3)	<i>p</i> value	Overall	
Creation time	<i>3colour</i>	104.68 (40.00)	102.81 (40.00)	71.44 (32.00)	n.s.	93.60 (36.13)	
	<i>single-colour</i>	115.37 (31.00)	95.96 (31.50)	92.89 (30.00)	n.s.	102.02 (31.00)	
Keystrok es	<i>3colour</i>	18.38 (15.00)	22.14 (19.00)	23.42 (20.00)	.021	21.17 (18.00)	
	<i>single-colour</i>	16.12 (14.00)	18.59 (16.00)	17.54 (15.00)	n.s.	17.36 (14.00)	

Creation time. There was no significant difference in the creation time between the media conditions for either the *3colour* indicator ($H(2) = 1.35$, $p = .510$) or *single-colour* indicator condition ($H(2) = 0.47$, $p = .789$). There was also no significant difference in this measure between the two colour-schemes ($Z = -0.52$, $p = .605$). However, compared to the *baseline* condition ($M = 21.24$, $Mdn = 19.00$), there was a significant difference in the creation time between the *baseline* and *3colour* ($Z = -8.56$, $p < .001$) conditions, and between the *baseline* and *single-colour* ($Z = -7.11$, $p < .001$)

conditions: participants spent a significantly longer time creating passwords in the *3colour* and *single-colour* than in the *baseline* condition.

Keystrokes. There was a significant difference in the number of keystrokes used to create a password between the three media conditions in the *3colour* indicator ($H(2) = 7.71, p = .021$), but not the *single-colour* indicator ($H(2) = 2.95, p = .229$). With the *3colour* indicator, participants who created passwords in the *graphical* condition used significantly fewer keystrokes than those in the *textual* and *graphical&textual* conditions. The pairwise comparison confirmed this difference (see Table 8.16). There was also a significant difference in the number of keystrokes between the two colour-schemes ($Z = -3.91, p < .001$). Namely, participants used significantly more keystrokes in the *3colour* than the *single-colour* condition. Comparing the two colour-schemes to the *baseline* condition ($M = 16.24, Mdn = 14.00$), there was a significant difference in the keystrokes between the *baseline* and *3colour* ($Z = -4.91, p < .001$), but not between the *baseline* and *single-colour* ($Z = -1.56, p = .118$): participants performed significantly more keystrokes in the *3colour* than in the *baseline* condition.

Table 8.16 Pairwise comparisons of keystrokes measures between media conditions for indicator

		<i>graphical</i>	<i>textual</i>	<i>graphical&textual</i>
Keystrokes	<i>3colour</i>	<i>graphical</i>	-	-24.14*
		<i>textual</i>	-	-28.38*
		<i>graphical&textual</i>		-4.25
				-

Note. * denotes a significant result in pairwise comparison, $p < .05$.

8.3.3.1.1.2 User Satisfaction Measures

Table 8.17 summarises the between-participants (in both colour-schemes) results for the six user satisfaction measures, and Table 8.18 illustrates the pairwise comparison results between the three media conditions (in both colour-schemes) for these measures.

Ease of use. There was no significant difference in the ratings of ease of use between the media conditions for either the *3colour* indicator ($H(2) = 1.00, p = .606$) or the *single-colour* indicator condition ($H(2) = 1.14, p = .565$). In addition, there was no significant difference in the ratings of this measure between the two colour-schemes

($Z = -0.80$, $p = .422$). Comparing the two colour-schemes to the *baseline* condition ($M = 2.04$, $Mdn = 1.00$), there was a significant difference in the ratings of ease of use between the *baseline* and *3colour* ($Z = -5.21$, $p < .001$), and between the *baseline* and *single-colour* ($Z = -4.17$, $p < .001$) conditions: participants rated the ease of use of the PCSs higher in the *3colour* and *single-colour* conditions than in the *baseline* condition.

Table 8.17 Mean (median) ratings of ease of use, annoyingness, helpfulness, clarity, amount of detail, and participants' confidence measures for the six indicator conditions

	Colour-scheme	Media			<i>p</i> value	Overall
		<i>graphical</i> (Group 1)	<i>textual</i> (Group 2)	<i>graphical&textual</i> (Group 3)		
Ease of use	<i>3colour</i>	2.24 (2.00)	2.40 (2.00)	2.46 (2.00)	n.s.	2.36 (2.00)
	<i>single-colour</i>	2.26 (2.00)	2.25 (2.00)	2.46 (2.00)	n.s.	2.32 (2.00)
Annoyingness	<i>3colour</i>	2.84 (3.00)	2.77 (3.00)	2.85 (3.00)	n.s.	2.82 (3.00)
	<i>single-colour</i>	2.94 (3.00)	2.60 (2.00)	2.91 (3.00)	n.s.	2.82 (3.00)
Helpfulness	<i>3colour</i>	2.72 (3.00)	2.77 (3.00)	2.99 (3.00)	n.s.	2.82 (3.00)
	<i>single-colour</i>	2.60 (2.00)	2.95 (3.00)	2.85 (3.00)	.028	2.79 (3.00)
Clarity	<i>3colour</i>	3.44 (3.00)	3.86 (4.00)	3.94 (4.00)	.009	3.73 (4.00)
	<i>single-colour</i>	3.29 (3.00)	3.72 (4.00)	3.84 (4.00)	.005	3.60 (4.00)
Amount of detail	<i>3colour</i>	1.99 (2.00)	1.90 (2.00)	2.01 (2.00)	n.s.	1.97 (2.00)
	<i>single-colour</i>	1.88 (2.00)	1.86 (2.00)	1.95 (2.00)	n.s.	1.89 (2.00)
Confidence	<i>3colour</i>	3.25 (3.00)	3.59 (4.00)	3.70 (4.00)	.027	3.50 (4.00)
	<i>single-colour</i>	3.02 (3.00)	3.81 (4.00)	3.63 (4.00)	.000	3.47 (3.00)

Annoyingness. There was no significant difference in the ratings of annoyingness between the media conditions for either the *3colour* indicator ($H(2) = 0.24$, $p = .885$) or the *single-colour* indicator ($H(2) = 3.68$, $p = .159$). There was also no significant difference in the ratings of this measure between the two colour-schemes ($Z = -0.04$, $p = .969$). However, comparing the two colour-schemes to the *baseline* condition ($M = 2.63$, $Mdn = 2.00$), there was a significant difference in the ratings of annoyingness between the *baseline* and *3colour* ($Z = -2.33$, $p = .020$), and between the *baseline* and *single-colour* ($Z = -3.01$, $p = .003$) conditions: the *3colour* and *single-colour* indicators were both perceived to be low in annoyingness.

Helpfulness. There was a significant difference in the ratings of helpfulness between the media conditions for the *single-colour* indicator ($H(2) = 7.14$, $p = .028$), but not for the *3colour* indicator ($H(2) = 4.07$, $p = .131$). For the *single-colour* indicator, the

ratings of helpfulness in the *textual* and *graphical&textual* conditions were significantly higher than in the *graphical* condition. The pairwise comparison confirmed this difference (see Table 8.18). However, there was no significant difference in the ratings of helpfulness between the two colour-schemes ($Z = -0.59$, $p = .556$). On the other hand, when the two colour-schemes were compared to the *baseline* condition ($M = 2.65$, $Mdn = 3.00$), there was a significant difference in the ratings of helpfulness between the *baseline* and *3colour* ($Z = -2.60$, $p = .009$), and between the *baseline* and *single-colour* ($Z = -2.16$, $p = .031$) conditions: participants perceived the *3colour* and *single-colour* conditions to be significantly more helpful than the *baseline* condition.

Table 8.18 Pairwise comparisons of helpfulness, clarity, and confidence measures between different media conditions for indicator

			<i>graphical</i>	<i>textual</i>	<i>graphical&textual</i>
Helpfulness	<i>single-colour</i>	<i>graphical</i>	-	-26.46*	-20.84*
		<i>textual</i>		-	-5.63
		<i>graphical&textual</i>			-
Clarity	<i>3colour</i>	<i>graphical</i>	-	-26.12*	-30.64*
		<i>textual</i>		-	-4.53
		<i>graphical&textual</i>			-
	<i>single-colour</i>	<i>graphical</i>	-	-25.10*	-34.10*
		<i>textual</i>		-	-9.00
		<i>graphical&textual</i>			-
Confidence	<i>3colour</i>	<i>graphical</i>	-	-22.65*	-27.14*
		<i>textual</i>		-	-4.49
		<i>graphical&textual</i>			-
	<i>single-colour</i>	<i>graphical</i>	-	-48.24*	-36.95*
		<i>textual</i>		-	11.28
		<i>graphical&textual</i>			-

Note. * denotes a significant result in pairwise comparison, $p < .05$.

Clarity. There was a significant difference in the ratings of clarity between the media conditions for both the *3colour* indicator ($H(2) = 9.47$, $p = .009$) and the *single-colour* indicator condition ($H(2) = 10.69$, $p = .005$): for both colour-schemes, the ratings of clarity in the *textual* and *graphical&textual* conditions were significantly higher than in the *graphical* condition. The pairwise comparison confirmed this difference (see Table 8.18). There was also a significant difference in the ratings of clarity between the two colour-schemes ($Z = -2.15$, $p = .032$): the clarity of the *3colour* indicator was perceived to be significantly higher than that of the *single-colour* indicator. Comparing the two colour-schemes to the *baseline* condition ($M = 3.68$, $Mdn = 4.00$), there was

no significant difference in the ratings of clarity between the *baseline* and *3colour* ($Z = -0.86, p = .388$), and between the *baseline* and *single-colour* ($Z = -1.25, p = .211$) conditions.

Amount of detail. There was no significant difference in the ratings of amount of detail between the media conditions for either the *3colour* indicator ($H(2) = 0.80, p = .672$) or the *single-colour* indicator condition ($H(2) = 0.591, p = .744$). There was also no significant difference in the ratings of this measure between the two colour-schemes ($Z = -1.80, p = .072$). Comparing the two colour-schemes to the *baseline* condition ($M = 1.68, Mdn = 2.00$), there was a significant difference in the ratings of amount of detail between the *baseline* and *3colour* ($Z = -5.48, p < .001$), and between the *baseline* and *single-colour* ($Z = -4.41, p < .001$) conditions: participants rated the amount of detail provided in the *3colour* and *single-colour* conditions significantly higher than in the *baseline* condition.

Confidence. There was a significant difference in the ratings of confidence between the media conditions for both the *3colour* indicator ($H(2) = 7.22, p = .027$) and the *single-colour* indicator conditions ($H(2) = 21.78, p < .001$). For both colour-schemes, the ratings of confidence in the *textual* and *graphical&textual* conditions were significantly higher than in the *graphical* condition. In contrast, the ratings of this measure in the *textual* condition were significantly higher than in the other media conditions for the *single-colour* indicator. The pairwise comparison confirmed this difference (see Table 8.18), but showed no significant difference between the *textual* and *graphical&textual* conditions. Furthermore, there was no significant difference in the ratings of confidence between the two colour-schemes ($Z = -0.49, p = .622$). When the two colour-schemes were compared to the *baseline* condition ($M = 3.23, Mdn = 3.00$), however, there was a significant difference in the ratings of confidence between the *baseline* and *3colour* ($Z = -3.39, p = .001$), and between the *baseline* and *single-colour* ($Z = -2.98, p = .003$) conditions. Participants felt more confident when the *3colour* and *single-colour* indicators were provided than they did in the *baseline* condition.

8.3.3.1.2 Strength of Password

8.3.3.1.2.1 Password Characteristics

Table 8.19 summarises the between-participants (in both colour-schemes) results for the password characteristics measures, and Table 8.20 illustrates the pairwise comparison results between the three media conditions (in both colour-schemes) for the password characteristics measures.

Password length. There was a significant difference in the total number of characters in passwords between the three media conditions in the *3colour* indicator ($H(2) = 8.97, p = .011$), but not in the *single-colour* indicator ($H(2) = 4.78, p = .092$) condition. In the *3colour* indicator condition, passwords created in the *graphical&textual* condition were longer than in the other two conditions. Furthermore, the pairwise comparison showed a significant difference between the *graphical&textual* and *graphical* conditions, but not the *textual* condition (see Table 8.20). In addition, there was a significant difference in the password length between the two colour-schemes ($Z = -2.65, p = .008$): passwords created with the *3colour* indicator were significantly longer than with the *single-colour* indicator. Comparing the two colour-schemes to the *baseline* condition ($M = 10.76, Mdn = 10.00$), there was a significant difference in the password length between the *baseline* and *3colour* ($Z = -5.55, p < .001$), and between the *baseline* and *single-colour* ($Z = -3.14, p = .002$) conditions: participants created significantly longer passwords in the *3colour* and *single-colour* conditions than in the *baseline* condition.

Number of digits. There was a significant difference in the number of digits used in passwords between the three media conditions for both the *3colour* indicator ($H(2) = 13.89, p = .001$) and the *single-colour* indicator ($H(2) = 6.14, p = .046$): in both colour-scheme conditions, passwords created in the *graphical&textual* condition contained more digits than in the other two conditions. In addition, the pairwise comparison showed a significant difference between the *graphical&textual* and the other two media conditions for the *3colour* indicator, whereas there was only a significant difference between the *graphical&textual* and *textual* conditions for the *single-colour*

indicator (see Table 8.20). However, there was no significant difference in the number of digits between the two colour-scheme conditions ($Z = -0.46$, $p = .647$). When the two colour-schemes were compared to the *baseline* condition ($M = 2.68$, $Mdn = 3.00$), there was no significant difference in the number of digits used in the passwords between the *baseline* and *3colour* ($Z = -1.33$, $p = .183$), and between the *baseline* and *single-colour* ($Z = -1.34$, $p = .181$) conditions.

Table 8.19 Mean (median) password length, number of digits, number of uppercase letters, number of lowercase letters, number of symbols, and strength score measures for the six indicator conditions

	Colour-scheme	Media			<i>p</i> value	Overall
		<i>graphical</i> (Group 1)	<i>textual</i> (Group 2)	<i>graphical&textual</i> (Group 3)		
Password length	<i>3colour</i>	11.17 (11.00)	11.95 (11.00)	12.59 (12.00)	.011	11.87 (12.00)
	<i>single-colour</i>	10.92 (11.00)	11.57 (12.00)	11.59 (11.00)	n.s.	11.34 (11.00)
Number of digits	<i>3colour</i>	2.40 (2.00)	2.83 (3.00)	3.43 (4.00)	.001	2.87 (3.00)
	<i>single-colour</i>	2.52 (3.00)	2.88 (3.00)	3.09 (4.00)	.046	2.81 (3.00)
Number of uppercase	<i>3colour</i>	1.07 (1.00)	0.90 (1.00)	1.30 (1.00)	n.s.	1.08 (1.00)
	<i>single-colour</i>	0.91 (1.00)	0.79 (1.00)	1.09 (1.00)	n.s.	0.93 (1.00)
Number of lowercase	<i>3colour</i>	6.57 (6.00)	6.73 (6.00)	6.60 (7.00)	n.s.	6.63 (6.00)
	<i>single-colour</i>	6.46 (6.00)	6.22 (6.00)	6.34 (6.00)	n.s.	6.34 (6.00)
Number of symbols	<i>3colour</i>	0.47 (0.00)	0.83 (1.00)	0.67 (1.00)	.021	0.65 (0.00)
	<i>single-colour</i>	0.35 (0.00)	0.69 (1.00)	0.70 (1.00)	.000	0.57 (0.00)
Strength score	<i>3colour</i>	64.91 (65.70)	70.03 (67.20)	75.46 (73.30)	.004	69.89 (71.50)
	<i>single-colour</i>	62.82 (59.13)	69.01 (65.70)	70.44 (71.45)	.011	67.22 (65.70)

Number of uppercase letters. There was no significant difference in the number of uppercase letters used in passwords between the three media conditions for either the *3colour* indicator ($H(2) = 3.25$, $p = .197$) or the *single-colour* indicator ($H(2) = 1.82$, $p = .402$). However, there was a significant difference in this measure between the two colour-schemes ($Z = -2.44$, $p = .015$): passwords created with the *3colour* indicator had significantly more uppercase letters than those created in the *single-colour* indicator. Comparing the two colour-schemes to the *baseline* condition ($M = 0.92$,

$Mdn = 1.00$), there was a significant difference in the number of uppercase letters in the passwords between the *baseline* and *3colour* ($Z = -2.35$, $p = .019$), but not between the *baseline* and *single-colour* ($Z = 0.00$, $p = 1.00$) conditions: passwords created using *3colour* indicator had significantly more uppercase letters than in the *baseline* condition.

Table 8.20 Pairwise comparisons of password length, number of digits, number of symbols, and strength score measures between media conditions for indicator

			<i>graphical</i>	<i>textual</i>	<i>graphical&textual</i>
Password length	<i>3colour</i>	<i>graphical</i>	-	-19.28	-33.40*
		<i>textual</i>		-	-14.11
		<i>graphical&textual</i>			-
Number of digits	<i>3colour</i>	<i>graphical</i>	-	-18.65	-41.45*
		<i>textual</i>		-	-22.80*
		<i>graphical&textual</i>			-
	<i>single-colour</i>	<i>graphical</i>	-	-18.03	-26.81*
		<i>textual</i>		-	-8.78
		<i>graphical&textual</i>			-
Number of symbols	<i>3colour</i>	<i>graphical</i>	-	-18.23	-27.96*
		<i>textual</i>		-	9.72
		<i>graphical&textual</i>			-
	<i>single-colour</i>	<i>graphical</i>	-	-34.22*	-34.25*
		<i>textual</i>		-	-0.03
		<i>graphical&textual</i>			-
Strength score	<i>3colour</i>	<i>graphical</i>	-	-18.61	-37.11*
		<i>textual</i>		-	-18.50
		<i>graphical&textual</i>			-
	<i>single-colour</i>	<i>graphical</i>	-	-25.32*	-31.60*
		<i>textual</i>		-	-6.28
		<i>graphical&textual</i>			-

Note. * denotes a significant result in pairwise comparison, $p < .05$.

Number of lowercase letters. There was no significant difference in the number of lowercase letters used in passwords between the three media conditions for either the *3colour* indicator ($H(2) = 0.13$, $p = .939$) or the *single-colour* indicator ($H(2) = 0.02$, $p = .989$). Moreover, there was also no significant difference in this measure between the two colour-scheme conditions ($Z = -1.74$, $p = .082$). Comparing the two colour-schemes to the *baseline* condition ($M = 6.16$, $Mdn = 6.00$), however, there was a significant difference in the number of lowercase letters used in the passwords between the *baseline* and *3colour* ($Z = -2.25$, $p = .025$), but not between the *baseline*

and *single-colour* ($Z = -0.42$, $p = .673$) conditions: passwords created using the *3colour* indicator had significantly more lowercase letters than in the *baseline* condition.

Number of symbols. There was a significant difference in the number of symbols used in passwords between the three media conditions for both the *3colour* indicator ($H(2) = 7.72$, $p = .021$) and the *single-colour* indicator ($H(2) = 15.29$, $p < .001$): for the *3colour* indicator, passwords created in the *textual* condition contained significantly more symbols, whereas for the *single-colour* indicator, this was the case for passwords created in the *graphical&textual* condition. The pairwise comparison showed a significant difference between only the *graphical&textual* and *graphical* conditions for the *3colour* indicator, and between the *graphical&textual* and the other two media conditions for the *single-colour* indicator (see Table 8.20). However, there was no significant difference in the number of symbols between the two colour-schemes ($Z = -1.82$, $p = .071$). On the other hand, when the two colour-schemes were compared to the *baseline* condition ($M = 0.38$, $Mdn = 0.00$), there was a significant difference in the number of symbols used in passwords between the *baseline* and *3colour* ($Z = -5.50$, $p < .001$), and between the *baseline* and *single-colour* ($Z = -4.74$, $p < .001$) conditions. Participants included significantly more symbols in their passwords in the *3colour* and *single-colour* conditions than in the *baseline* condition.

Number of password character classes. Regarding the number of different character classes in the passwords, there were significant differences in all six conditions: *3colour-graphical* (Group 1: $\chi^2(3) = 21.02$, $p < .001$), *3colour-textual* (Group 2: $\chi^2(3) = 12.76$, $p = .005$), *3colour-graphical&textual* (Group 3: $\chi^2(3) = 19.10$, $p < .001$), *single-graphical* (Group 1: $\chi^2(3) = 17.93$, $p < .001$), *single-textual* (Group 2: $\chi^2(3) = 13.72$, $p = .003$), and *single-graphical&textual* (Group 3: $\chi^2(3) = 17.91$, $p < .001$). Figure 8.25 shows the distribution of this measure across the six indicator conditions.

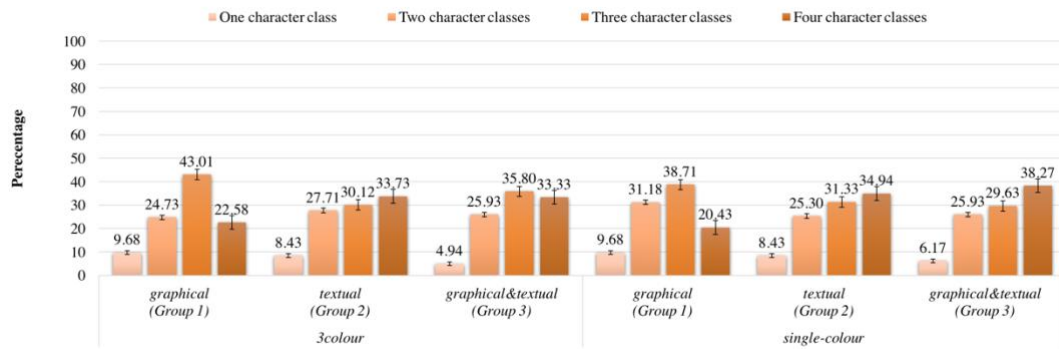


Figure 8.25 Percentage of password character classes across six indicator conditions

For the *3colour* indicator, the *textual* condition had the highest percentage of passwords using all four possible character classes (33.70%), whereas for the *single-colour* indicator, both the *textual* (34.90%) and *graphical&textual* (38.30%) conditions had the highest percentage of passwords using all four character classes. However, there was no significant difference in this measure between the two colour-schemes ($Z = -0.21$, $p = .833$). On the other hand, when the two colour-schemes were compared to the *baseline* condition, there was a significant difference in the number of character classes between the *baseline* and *3colour* ($Z = -4.76$, $p < .001$), and between the *baseline* and *single-colour* ($Z = -4.54$, $p < .001$) conditions. Namely, in both colour-schemes, most passwords had between three and four character classes: for three, *3colour* (36.58%) and *single-colour* (33.46%); and for four, *3colour* (29.57%) and *single-colour* (30.74%). On the other hand, in the *baseline* condition, most of passwords included two (30.74%) or three (35.41%) character classes.

Password strength score. There was a significant difference in password strength scores between the three media conditions for the *3colour* indicator ($H(2) = 10.81$, $p = .004$) and the *single-colour* indicator ($H(2) = 8.97$, $p = .011$): in both colour-scheme conditions, passwords created with the *graphical&textual* indicator were stronger than in the other two conditions. The pairwise comparison showed a significant difference only between the *graphical&textual* and *graphical* conditions in the two colour-schemes (see Table 8.20).

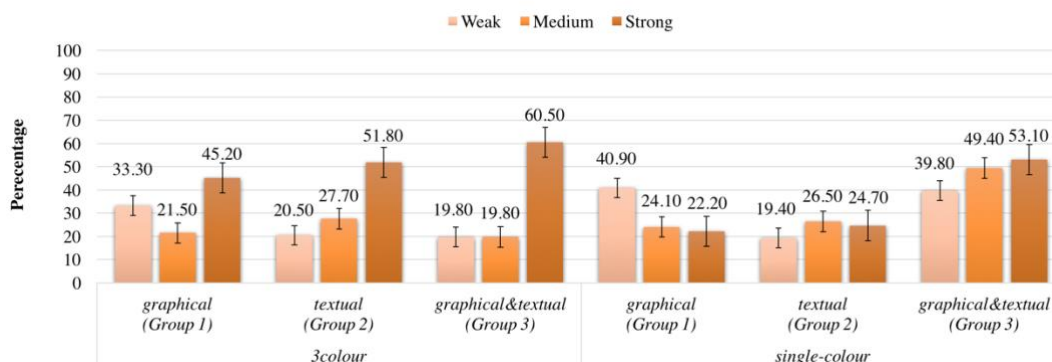


Figure 8.26 Percentage of password strength levels across the six indicator conditions

Figure 8.26 shows the distribution of password strength levels across the different conditions. In addition, there was a significant difference in the password strength scores between the two colour-schemes ($Z = -2.02, p = .044$): passwords created in the *3colour* condition were significantly stronger than in the *single-colour* condition.

8.3.3.1.2.2 Password Guessability

Since participants were not forced to follow any password policy, all 257 passwords were examined to check their guessability using five cracking approaches (based on Ur et al. (2015)). For the *3colour* indicator, 150 (58.37%) passwords were not guessable, whereas 107 (41.63%) were guessable by at least one of the five approaches. For the *single-colour* indicator, 126 (49.03%) passwords were not guessable, while 131 (50.97%) were guessed. The distribution of the password guessability across the six indicator conditions is shown in Figure 8.27.

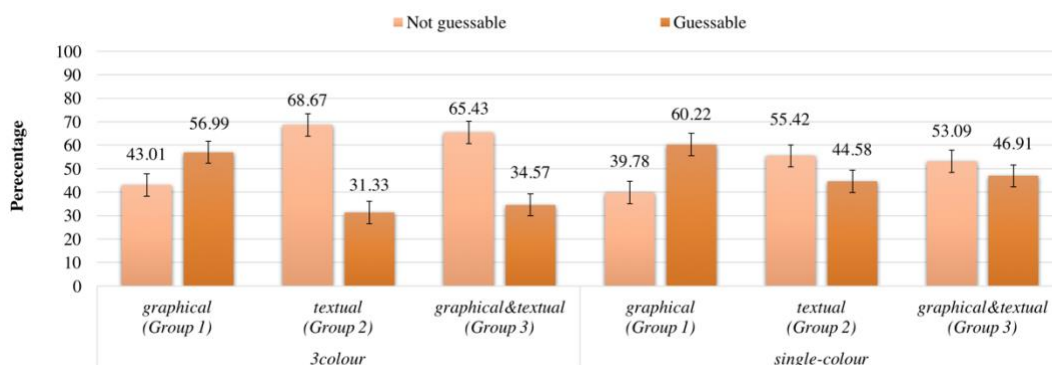


Figure 8.27 Percentage of password guessability across the six indicator conditions

Regarding the six indicator conditions, there was a significant difference in the number of guessable and non-guessable passwords in two conditions for the *3colour* indicator, and in only one condition for the *single-colour* indicator: *3colour-textual* (Group 2: $x^2(1) = 1.82$, $p = .001$), *3colour-graphical&textual* (Group 3: $x^2(1) = 7.72$, $p = .005$), and *single-graphical* (Group 1: $x^2(1) = 3.88$, $p = .049$). However, there was no significant difference in the other conditions: *3colour-graphical* (Group 1: $x^2(1) = 1.82$, $p = .178$), *single-textual* (Group 2: $x^2(1) = 0.98$, $p = .323$), and *single-graphical&textual* (Group 3: $x^2(1) = 0.31$, $p = .579$). For the *3colour* indicator, the *3colour-textual* condition had the highest percentage of non-guessable passwords (68.67%). On the other hand, for the *single-colour* indicator, the *single-graphical* condition had the high percentage of passwords (39.78%) that were guessable.

8.3.3.2 Password Recall

The following section presents the results regarding the usability of PCS when users recall passwords.

8.3.3.2.1 Efficiency Measures

Recall time. There was no significant difference in the recall time between the media conditions for either the *3colour* indicator ($H(2) = 0.44$, $p = .804$) or the *single-colour* indicator condition ($H(2) = 1.73$, $p = .422$). Table 8.21 shows the mean recall time for the six indicator conditions. There was also no significant difference in this measure between the two colour-schemes ($Z = -0.33$, $p = .741$).

Table 8.21 Mean (median) recall time measure for the six indicator conditions

	Colour-scheme	Media			<i>p</i> value	Overall
		<i>graphical</i> (Group 1)	<i>textual</i> (Group 2)	<i>graphical&textual</i> (Group 3)		
Recall time	<i>3colour</i>	19.22 (16.00)	20.08 (16.00)	20.87 (19.00)	n.s.	19.99 (17.00)
	<i>single-colour</i>	18.18 (17.00)	20.19 (16.00)	21.65 (19.00)	n.s.	19.86 (17.00)

However, comparing the two colour-schemes to the *baseline* condition ($M = 39.80$, $Mdn = 38.00$), there was a significant difference in the recall time between the *baseline* and *3colour* ($Z = -10.84$, $p < .001$), and between the *baseline* and *single-colour* ($Z = -10.32$, $p < .001$) conditions. Participants spent less time recalling the passwords

created in the *3colour* and *single-colour* conditions than those created in the *baseline* condition.

8.3.3.2.2 Effectiveness Measures

Accuracy. There were significant differences in the accuracy of recalling passwords in two media conditions for the *3colour* indicator, and in all media conditions for the *single-colour* indicator: *3colour-graphical* (Group 1: $x^2(1) = 8.14$, $p = .004$), *3colour-graphical&textual* (Group 3: $x^2(1) = 4.41$, $p = .036$), *single-graphical* (Group 1: $x^2(1) = 8.14$, $p = .004$), *single-textual* (Group 2: $x^2(1) = 11.07$, $p = .001$), and *single-graphical&textual* (Group 3: $x^2(1) = 8.65$, $p = .003$). In contrast, there was no significant difference in the *3colour-textual* (Group 2: $x^2(1) = 2.77$, $p = .096$) condition. Figure 8.28 shows the percentage of accuracy in password recall across all six indicator conditions.

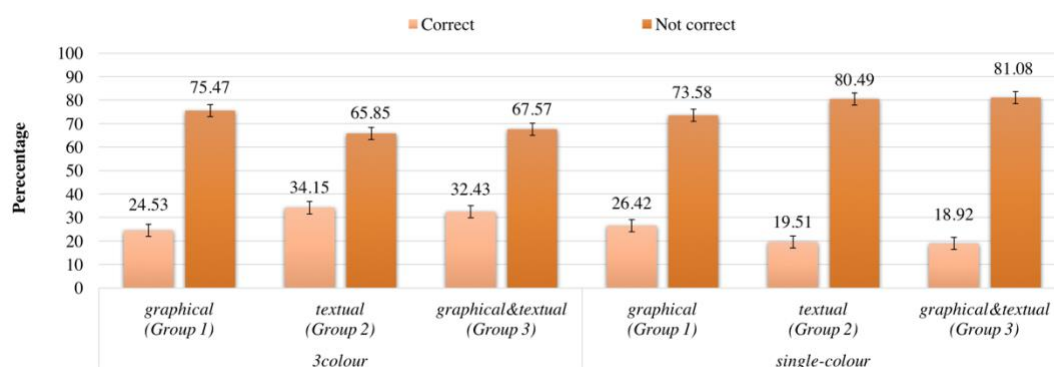


Figure 8.28 Percentage of accuracy in recalling passwords across the six indicator conditions

For the *3colour* indicator, the percentage of successful recall was higher in the *graphical&textual* (32.43%) than in the *graphical* condition (24.53%), whereas for the *single-colour* indicator, the *graphical* condition (26.42%) had the highest successful recall rate. However, there was no significant difference in the accuracy level between the two colour-schemes ($Z = -1.67$, $p = .095$). On the other hand, comparing the two colour-schemes to the *baseline* condition, there was a significant difference in the accuracy level between the *baseline* and *3colour* ($Z = -2.21$, $p = .027$), and between the *baseline* and *single-colour* ($Z = -3.36$, $p = .001$) conditions. Namely, the accuracy of recalling passwords correctly in the *3colour* (35.12%) and *single-*

colour indicator (29.76%) conditions was lower than in the *baseline* (43.45%) condition.

8.3.3.2.3 User Satisfaction Measures

Confidence. Table 8.22 summarises the between-participants (in both colour-scheme conditions) results for the confidence measure. There was no significant difference in the ratings of confidence in recalling the correct passwords between the media conditions for either the *3colour* indicator ($H(2) = 1.49, p = .476$) or the *single-colour* indicator condition ($H(2) = 1.74, p = .418$). There was also no significant difference in this measure between the two colour-schemes ($Z = -1.17, p = .244$). When the two colour-schemes were compared to the *baseline* condition ($M = 3.01, Mdn = 3.00$), however, there was a significant difference in the ratings of confidence between the *baseline* and *3colour* ($Z = -3.75, p < .001$), and between the *baseline* and *single-colour* ($Z = -3.49, p < .001$) conditions: participants felt significantly less confident in recalling the correct passwords in both the *3colour* and *single-colour* indicator conditions than they did in the *baseline* condition.

Table 8.22 Mean (median) ratings of confidence measure for the six indicator conditions

	Colour-scheme	Media			<i>p</i> value	Overall
		<i>graphical</i> (Group 1)	<i>textual</i> (Group 2)	<i>graphical&textual</i> (Group 3)		
Confidence	<i>3colour</i>	2.51 (2.00)	2.71 (3.00)	2.76 (3.00)	n.s.	2.65 (3.00)
	<i>single-colour</i>	2.57 (2.00)	2.81 (3.00)	2.86 (3.00)	n.s.	2.73 (3.00)

8.3.3.3 Summary

For the password creation, the three media presentations of the password strength indicator differed significantly in the number of keystrokes (only *3colour*), perceived helpfulness (only *single-colour*), perceived clarity, participants' confidence, password length (only *3colour*), number of digits, number of symbols, number of character classes, and password strength score. Participants used significantly less effort in creating passwords with the *3colour-graphical* indicator. Overall, the ratings of helpfulness, clarity, and participants' confidence were higher in the *textual* and

graphical&textual conditions than in the *graphical* condition for both colour-schemes. Furthermore, the passwords created in the *graphical&textual* condition were longer and had higher strength scores than in the other conditions. Furthermore, the *graphical&textual* condition encouraged participants to include more digits and symbols, and four character classes. Overall, the percentage of non-guessable passwords in the *3colour* indicator was higher than that in the *single-colour* indicator condition. Specifically, the *3colour-textual* and *3colour-graphical&textual* conditions had the highest percentage of non-guessable passwords.

The usability of PCSs and password strength differed significantly between the *3colour* and *single-colour* conditions in the number of keystrokes, perceived clarity, password length, number of uppercase letters, and finally password strength score. Providing a *3colour* indicator made participants try more keystrokes, and resulted in longer passwords that included more uppercase letters and scored very high in terms of strength compared to the *single-colour* indicator. Furthermore, the *3colour* indicator made the password creation process clearer than the *single-colour* indicator did.

Regardless of the colour-scheme indicator conditions, the usability of PCSs and passwords strength between the *indicator* and *baseline* conditions differed significantly on the efficiency of the PCSs (except keystrokes in the *single-colour*), level of user satisfaction (except clarity) on the PCSs and password characteristics (except number of digits, number of uppercase letters and lowercase letters in the *single-colour*). Providing an *indicator* regardless of the colour-scheme made the password creation process less efficient, but it did improve the level of user satisfaction and the password characteristics.

For the password recall, the three media presentation conditions for both colour-schemes did not differ significantly in the recall time and participants' confidence. However, the level of accuracy differed significantly in media presentation conditions (except *3colour-textual*): successful recall rates were low compared to the unsuccessful ones. There was also no significant difference in the recall time, accuracy level, and participants' confidence between the *3colour* and *single-colour* indicators. However, a significant difference was found between the *3colour* and

baseline conditions, and between the *single-colour* and *baseline* conditions for these measures. Participants spent significantly less time recalling passwords in the *3colour* and *single-colour* conditions, but they did not recall correct passwords and felt less confident.

8.3.4 Users' Common Password Creation and Recall Practices

Participants reported that they had on average around 19.98 (standard deviation = 25.41) password-protected accounts, and approximately 11.90 (standard deviation = 17.04) passwords. However, the majority reported using the same password (181, 70.43%) or slightly different passwords (184, 71.60%) for multiple accounts. When asked about their behaviour in this study, very few on average reported creating their passwords based on reused ones (24, 9.34%) or modified ones (43, 16.73%). In contrast, the majority of participants on average reported creating entirely new passwords (190, 73.93%). This might be because they were instructed to try their best to create a new password for this study.

Regarding password change frequency, many participants (109, 42.41%) reported changing their passwords every three to six months, while very few (17, 6.61%) never did so. Many participants described themselves as being very knowledgeable (92, 35.80%) and moderately knowledgeable (81, 31.52%) of what makes a secure password. Participants in this study had a similar perception of what makes a secure password as those in Studies 4 and 5 (see Chapter 6 and 7) in terms of having a combination of different character classes, not being based on personal information, and containing a large number of characters. In addition, very few participants mentioned that uniqueness is what makes a secure password. Interestingly, none of the participants in this study mentioned the use of a password management system as a criterion for making secure passwords. Several participants (91, 35.41%) felt very confident about the strength of their most complicated password, while only one (1, 0.39%) did not feel confident at all.

Regarding password creation instructions, most participants (121, 47.08%) rated their frequency of reading the instructions when creating a new password as being 'always'.

However, there were some circumstances under which they did not do so; some were related to the participants, and some to the instructions themselves. These circumstances were almost the same as those found in Studies 4 and 5 (see Chapter 6 and 7). The circumstances that related to participants managing to create a new password on the first attempt, were familiarity with the website, and being in a hurry. The circumstances that related to the instructions were being too lengthy, being invisible in the PCS, or being associated with low-value accounts.

Few participants (36, 14.01%) reported having a negative experience during the password creation process. What participants explained about their frustration was in line with the findings of Study 4 (see chapter 6), as follows: (1) PCSs enforced a very strict password policy (e.g. including uppercase letters and symbols in the password) but were not associated with high-value account, and (2) it took a long time trying to comply with the password policy.

Regarding password management strategies, participants mentioned writing down passwords, using a password manager, reusing the same passwords for multiple accounts, modifying different variations of the same passwords, relying on their memory, or choosing passwords that were easy to remember. Furthermore, they reported different ways of keeping them safe when they chose to write them down, such as using notepads, sticky notes, or encrypted files on their computer. In total, 37 (22.02%) participants reported writing down their passwords they created for this study.

8.4 Discussion

This study aimed to investigate the individual effects of presenting the password policy, creation suggestion, and strength indicator to users in a PCS. It specifically examined the best timing at which to present the password policy and creation suggestion, and the best media and colour-scheme presentation of strength indicators. To this end, four timing of presentation conditions (*before-interaction*, *during-interaction*, *after-interaction* and *during&after-interaction*) were examined for the password policy and creation suggestion; and three media conditions (*graphical*,

textual, and *graphical&textual*) and two colour-scheme conditions (*3colour* and *single-colour*) were examined for strength indicator. The study consisted of two parts: (1) password creation and (2) password recall. A total of 257 users produced usable data in the first part, while 168 of these returned for the second one.

The first and second research questions (RQ1 and RQ2) concerned the password policy, the third and fourth (RQ3 and RQ4) the creation suggestion, and the fifth and sixth (RQ5 and RQ6) the strength indicator. Finally, the seventh and eighth (RQ7 and RQ8) compared the provision of a supporting feature to the baseline. All research questions examined both password creation (in terms of the usability of PCSs and password strength) and recall (in terms of the usability of PCSs). These research questions are answered according to the type of supporting feature, as follows: password policy in Section 8.4.1, password creation suggestion in Section 8.4.2, and password strength indicator in Section 8.4.3.

8.4.1 Password Policy

The present study found that the timing of presentation of password policy had an effect on the efficiency of PCSs and password guessability, but not on user satisfaction and password characteristics (except password length, number of character classes, and policy compliance) when users created passwords. On the other hand, the study did not find an effect of different timings of presentation of password policy on the efficiency of the PCSs and user satisfaction when users recalled passwords, but there was an effect on the effectiveness of the PCSs.

In terms of password creation, the four timings of presentation affected the usability of PCSs and the password strength differently. To some extent, both the *policy-before-interaction* and *policy-during-interaction* timings, and both the *policy-after-interaction* and *policy-during&after-interaction* timings had similar effects on the password creation when they were implemented in the PCS.

The results of this study indicated that using the *policy-before-interaction* and *policy-during-interaction* timings of presentation made the password creation process efficient in terms of the time and effort required to complete the task (see Table 8.4).

As a result, passwords created in these conditions were shorter than in the others (see Table 8.7). Furthermore, these conditions also had a lower percentage of passwords that included all four character classes compared to the other conditions (see Figure 8.12). However, if the PCSs did not enforce the given policy and check the created passwords, the *policy-during-interaction* presentation led to a considerable failure to comply with the policy, with almost half of the passwords failing to comply (see Figure 8.13). This might explain why participants who created passwords using the *during-interaction* presentation completed the task quickly and with less effort: they were not creating strong, policy-compliant passwords. On the other hand, although the *policy-during-interaction* presentation did not significantly encourage users to create compliant passwords, the *policy-before-interaction* presentation did, as the second highest percentage of compliant passwords were created in this condition.

Furthermore, the results showed that the *policy-after-interaction* and *policy-during&after-interaction* timings of presentation made the password creation process more time consuming. It seems that users in these two conditions thought about the constituents of passwords more carefully than those in the other two conditions: the passwords they created were longer, and these conditions had the highest percentages of passwords including all four character classes. A possible explanation of this result might be that users interpreted the policy statement at this stage as an error message that would not let them proceed with the creation process until they improved their passwords. In terms of policy compliance, the *policy-during&after-interaction* presentation had a higher percentage of compliant passwords than the *after-interaction* presentation. However, only the *policy-after-interaction* presentation required more effort, which was because the users in the *policy-during&after-interaction* presentation could make the necessary changes while entering their passwords, thereby saving them keystrokes.

Regarding password guessability, all four timings of presentation had a high percentage of non-guessable passwords compared to guessable ones (see Figure 8.14). This might be because the chosen policy was proven to create secure passwords (Shay et al., 2014).

It was somewhat unexpected that the results of this study revealed no effect on the level of user satisfaction with the PCSs between the four timing of presentation conditions for the policy (see Table 8.6). It is difficult to explain this result, but it may be related to the chosen policy itself and not to the design of the PCSs. Some participants expressed their frustration with the chosen policy. For example, one participant commented: *'12 characters seems a little excessive as it is more difficult to remember'*. Another wrote: *'This is the most annoying type because it requires so many characters and so many different kinds of characters. I can never remember passwords with this much detail and they end up making me feel less confident, not more.'* A stringent password policy was used in the study to make the password creation task effortful for the participants, but it may have been too onerous.

In terms of password recall, the four timings of presentation did not affect the recall time and participants' confidence (see Figure 8.15 and Figure 8.17). On the other hand, the level of accuracy differed significantly in all four timing of presentation: successful recall rates were low compared to the unsuccessful ones (see Figure 8.16). The *policy-before-interaction* presentation had the highest percentage of successful recall, while *policy-during-interaction* had the highest unsuccessful recall rate.

Another important finding was that providing a policy statement (regardless of the timing of presentation) affected the usability of the PCSs and the strength of the passwords that the users created, and also affected the usability when users recalled their passwords. In terms of the password creation, providing a policy statement negatively affected the efficiency of the PCSs compared to not providing one. However, it did also positively improve the level of user satisfaction (except annoyingness) and the password strength. Finally, in terms of the password recall, providing a policy statement negatively affected the usability of the PCSs, resulting in a low accuracy level and confidence rate in comparison to not providing such a statement.

8.4.2 Password Creation Suggestion

The present study found that the timing of presentation of the password creation suggestion had an effect on the efficiency of PCSs, but not on user satisfaction, password characteristics (except password length, number of character classes, suggestion compliance, and symbols provision), and password guessability when users created passwords. On the other hand, no effect of different timings was found on the efficiency of the PCSs and user satisfaction when users recalled passwords, whereas there was an impact on the effectiveness of the PCSs.

In terms of the password creation, the four timings of presentation affected the usability of PCS and the password strength differently. To some extent, it was difficult to find patterns of similarity or difference between the timings of presentation when they were implemented in the PCS.

The results of this study indicated that presenting the suggestion statement using either the *suggestion-before-interaction* or the *suggestion-during&after-interaction* presentation made the password creation quicker with little effort required (see Table 8.9). However, passwords created with the *suggestion-before-interaction* presentation were shorter compared to with the other three presentations. Furthermore, the *suggestion-before-interaction* condition had a low percentage of passwords including all four character classes, although the majority of passwords in this presentation were compliant with the given suggestion (see Figure 8.18 and Figure 8.19). Participants felt very confident creating passwords in both the *suggestion-during&after-interaction* and *suggestion-during-interaction* conditions (see Table 8.17); and both presentations had a high percentage of passwords including all four character classes.

However, concerns about the password predictability arise, since the results confirmed that users tended to use the examples given in the suggestion statement in their passwords, which might decrease their security. This occurred especially when these examples were presented in the *suggestion-during-interaction* or *suggestion-after-interaction* timing (see Figure 8.20). Furthermore, none of the four timings of presentation had a high percentage of non-guessable passwords, even though there was significant difference in the *suggestion-before-interaction* presentation (see

Figure 8.21). It seems possible that encouraging users to have only a combination of letters, symbols, and digits (without mentioning the length of passwords) was not enough to create strong passwords.

In terms of password recall, the four timings of presentation did not affect the recall time and participants' confidence (see Figure 8.22 and Figure 8.24). On the other hand, the level of accuracy differed significantly in two conditions (*suggestion-before-interaction* and *suggestion-during&after-interaction*): successful recall rates were low compared to the unsuccessful ones (see Figure 8.23). Furthermore, the *suggestion-during&after-interaction* presentation had the highest percentage of successful recall, while the *suggestion-before-interaction* presentation had the highest unsuccessful recall rate.

Another important finding was that providing a suggestion statement (regardless of the timing of presentation) affected the usability of the PCSs, but not the strength of the passwords users created. At the same time, it also affected the PCS usability when users recalled their passwords. These effects were similar to the effect of providing a policy statement during the password creation process. Furthermore, in terms of password creation, providing a suggestion statement negatively affected the efficiency of the PCSs compared to not providing one, yet it also improved the level of user satisfaction (except annoyingness and clarity). However, providing a suggestion statement did not have an effect on the password strength (except number of symbols and password character classes). For the number of symbols, the difference might be due to the effect of providing them in the suggestion statement. Finally, in terms of the password recall, providing a suggestion statement negatively affected the usability of the PCSs by resulting in a low accuracy level and confidence rate in comparison to not providing one.

8.4.3 Password Strength Indicator

The current study found that the media and colour-scheme presentation of password strength indicator affected the efficiency of PCSs (except creation time), password characteristics, and password guessability, but not the level of user satisfaction when

users created passwords. On the other hand, this study did not find an effect of different media and colour-schemes of the password strength indicator on the efficiency of the PCSs and user satisfaction when users recalled passwords, but did find an effect on the effectiveness of the PCSs.

In terms of password creation, the media and colour-scheme presentations had an interesting pattern of significant effects on the keystrokes, password length and characteristics, and perceived usability of the indicators. The *graphical&textual* indicator often produced stronger and more complex passwords, particularly in the *3colour* condition (see Table 8.19). It was also perceived as clearer and more helpful, and it made participants more confident (see Table 8.17). With regard to colour-scheme, passwords created with the *3colour* indicator were typically longer and stronger, and had more uppercase letters than those in the *single-colour* indicator condition. In addition, the *3colour* indicator had higher ratings on perceived clarity compared to the *single-colour* indicator. Overall, the percentage of non-guessable passwords in the *3colour* condition was higher than in the *single-colour* condition. Specifically, passwords created in the *3colour-textual* and *3colour-graphical&textual* conditions had the highest percentage of non-guessable passwords (see Figure 8.27).

These results have interesting implications for designing strength indicators. Providing a *graphical* indicator without explaining what the changes in the bar mean may result in weaker passwords and poor usability. In addition, using only one colour to distinguish between the strength levels may also result in weaker passwords and poor usability, whereas using the traffic light metaphor of green, amber, and red colours results in stronger passwords.

In terms of password recall, media and colour-scheme presentations of the password strength indicator did not affect the recall time and participants' confidence (see Table 8.21 and Table 8.22). However, the level of accuracy differed significantly in for both colour-scheme conditions: successful recall rates were low compared to the unsuccessful ones (see Figure 8.28).

Another important finding was that providing a password strength indicator differently affected the usability of the PCSs and the strength of the passwords users created

depending on the colour-scheme, but had the same effect for both colour-schemes when users recalled passwords. Specifically, for password creation, providing a *3colour* indicator affected the PCS usability and the strength of passwords (except number of digits), but providing a *single-colour* affected only the former. Moreover, they both negatively affected the efficiency of the PCSs compared to not providing them, yet both improved the level of user satisfaction (except clarity). For the password recall, providing a password strength indicator with either colour-scheme negatively affected the usability of the PCSs by resulting in a low accuracy level and confidence rate in comparison to not providing the strength indicator.

8.5 Conclusions

In general, the findings of this study suggest that different presentations of the supporting features affected the usability of the PCSs and the password strength differently when users created passwords. When presenting the policy statement, the timing had an effect on the efficiency of PCSs and password guessability, but not on the level of user satisfaction and password characteristics. Regarding the suggestion statement, the timing of presentation had an effect on the efficiency of PCSs, but not on the level of user satisfaction, password characteristics and password guessability. Finally, for the password strength indicator, the media and colour-scheme presentations had an effect on the efficiency of PCSs, password characteristics, and password guessability, but not on the level of user satisfaction when creating passwords.

One of the significant findings to emerge from this study is that the presence of supporting features is important to improve the usability of PCSs and the strength of passwords. In general, the use of a password policy statement, creation suggestion, and password strength indicator improves the perceived usability of PCSs. Furthermore, as expected, providing a password policy within PCSs also improves the strength of passwords.

However, these results must be considered in light of the limitations, some of them discussed in Study 4 (see Chapter 6). Firstly, the task in this study lacked ecological

validity since participants imagined a situation where they need to create passwords for their online bank account. We do not know whether their behaviour would be similar in the real situation, particularly as they might be quite stressed if their online bank account had been compromised. However, the fact that there were many significant differences between the conditions on multiple dependent variables suggests that participants were taking the scenario and the PCS seriously, as different versions of the PCS created different behaviours. In addition, the fact that nearly three quarters of participants (73.39%) reported creating an entirely new password suggested they did make an effort while doing the study. The second limitation was in the password recall task. Asking participants to make five passwords and then recall them all three days later may create confusion between the different passwords in their minds and may negatively affect the recall rates. Indeed, only about two thirds of participants (65.37%) returned for the recall task, in spite of the fact that they would have earned a further USD 0.70. In terms of the recall accuracy, there was a very poor recall rate which might be due to the following reasons. First, the participants were asked to remember five passwords in one setting. Second, there was no penalty for not recalling the password correctly. For future research, a bonus could be offered as an incentive for correctly recalling the password. An analysis was conducted on the passwords recalled for both confusion between the passwords and for accuracy based on the order in which they were created, but neither of these factors had a big effect on the results. Furthermore, 22% of participants who returned for the second part of the study reported writing down their passwords, which suggests that they behaved in the same way they would normally do when managing their passwords.

To conclude, these results suggest that more attention should be paid to improving the design of PCSs as whole interactive systems with respect to their supporting features. Poor design of these aspects could affect password strength in different ways. For instance, providing examples of symbols during the password creation process makes passwords more predictable. All in all, as current PCSs implement more than one supporting feature for users, next chapter (Study 7) discusses the combined effects of presenting these supporting features, taking into account the outcomes of the present study.

Chapter 9

The Combined Effects of Supporting Features on Password Creation and Recall – *Study 7*

9.1 Introduction

The findings from the previous study (Study 6) showed that different presentations of each supporting feature affect the usability of PCSs and the password strength differently. However, what is the effect of presenting more than one supporting feature in a PCS? Is it the sum of the effects of single features or do features interact with each other in more complex ways? The exploratory analysis (see Study 1, Chapter 3) revealed that this is quite a common situation, as more than 40% of current PCSs (43.44%) present at least two supporting PCS features to users during the password creation process. In addition, the author has not found any research that investigates either the effect of combining supporting features or the combinations that might best help users during the password creation process. During the user evaluation (Study 3), the author noticed that users easily became confused when they were offered both a password policy and a creation suggestion during the password creation process.

Therefore, the present study aims to examine the effects of combining these supporting features to users in a PCS. To this end, it used the best presentation of each supporting feature identified in Study 6 (see Section 8.3, Chapter 8) to design four combinations of supporting features. The study consisted of two parts: password creation (Part I) and password recall (Part II). In Part I, each participant was asked to create two passwords with and without a combination of supporting features. In Part II, participants were asked to recall their passwords three days later. A total of 220

participants from the MTurk completed Part I appropriately, but only 147 (66.82%) responded to Part II within 24 hours of the invitation being sent.

The following research questions are formulated to address this study's aim:

RQ1. Are there differences in PCS usability and password strength between different combinations of supporting features when users create passwords?

RQ2. Are there differences in PCS usability between different combinations of supporting features when users recall passwords?

RQ3. Are there differences in PCS usability and password strength between providing combined supporting features and not providing them (i.e. baseline) when users create passwords?

RQ4. Are there differences in PCS usability between providing combined supporting features and not providing them (i.e. baseline) when users recall passwords?

9.2 Method

9.2.1 Design

Table 9.1 presents the study design and conditions. This study used a mixed design with one between-participants factor and one within-participant factor.

Table 9.1 Study design and conditions

	provision of combined supporting features	
	<i>baseline</i>	<i>combination</i>
	types of combination	
Group 1	<i>baseline</i>	<i>policy&suggestion</i>
Group 2	<i>baseline</i>	<i>policy&indicator</i>
Group 3	<i>baseline</i>	<i>suggestion&indicator</i>
Group 4	<i>baseline</i>	<i>policy&suggestion&indicator</i>

The between-participants factor is the type of combination of supporting features. All possible combinations of the three supporting features were considered, which resulted in four conditions: *policy&suggestion*, *policy&indicator*,

suggestion&indicator, and *policy&suggestion&indicator*. The design of the presentation of each supporting feature was based on findings from Study 6 (see chapter 8), as follows: the *policy-before-interaction* presentation was used for the password *policy*, the *suggestion-during-interaction* presentation was employed for the password creation *suggestion*, and the *3colour-graphical&textual* presentation was the basis for the password strength *indicator*. The *policy-before-interaction* presentation was chosen as it required less time and effort from users to create passwords, produced high compliant password rates, and resulted in very successful recall rates. The *suggestion-during-interaction* presentation was selected as it required less time from users, produced long passwords that had the full character classes, and had a high perceived confidence level. The *3colour-graphical&textual* presentation was chosen as it had high level of perceived user satisfaction and it also produced stronger passwords that were long and had four character classes.

The within-participants factor is the provision of the combined supporting features. It has two conditions: one without a combined supporting features (*baseline*) and one with a combined supporting features (*combination*).

As noted previously, this study consisted of two parts. In Part I, each participant was asked to create two passwords: one in the baseline condition and one in one of the four combination conditions (i.e. *policy&suggestion*, *policy&indicator*, *suggestion&indicator*, and *policy&suggestion&indicator*). There were thus four groups of participants, one for each of the four combination condition types. Each participant was randomly assigned to one of these groups. In Part II, all participants were asked to recall their two passwords three days later after completing Part I.

The dependent measures in the present study were similar to those in Studies 4 and 6 (see Section 6.2.1 in Chapter 6 and Section 8.2.1 in Chapter 8, respectively). Part I included two groups of dependent measures: those related to the usability of the PCS and those related to the strength of the password. Efficiency and user satisfaction measures were included to assess the usability of the PCSs. The time and number of keystrokes used to create a password were included as efficiency measures, whereas, participants' ratings (using a five-point Likert item) regarding ease of use,

annoyingness, helpfulness, clarity, amount of detail, and confidence in using the PCSs were included as user satisfaction measures. An optional open-ended question also gave participants the chance to explain their ratings. In terms of password strength, password characteristics and password guessability measures were used. The password characteristics included the following measures: password length, number of digits, number of uppercase letters, number of lowercase letters, number of symbols, and number of character classes utilised in the passwords. There were also measures that related to particular supporting features; one additional measure for the password policy, two for the password creation suggestion, and finally one for the password strength indicator. A policy compliance measure was used for the password policy. A suggestion compliance and a symbols provision measures for the password creation suggestion. A password strength score (based on Egelman et al., 2013) for the password strength indicator. The password guessability measure checked the created passwords in terms of the ability to guess each password over five cracking approaches (based on Ur et al., 2015).

Part II included three dependent measures of the usability of the PCS. These measures were the participant's time to recall a password, accuracy in password recall, and confidence in recalling his or her password correctly.

9.2.2 Participants

A total of 270 participants from MTurk took part in the study; 50 entries were excluded because the responses were incomplete (35 entries) or included creating identical passwords for the two conditions (15 entries). This resulted in 220 participants being included in the analysis. Participants were randomly assigned to one of the four conditions: this resulted in *Group 1* (57 participants), *Group 2* (56 participants), *Group 3* (59 participants), and *Group 4* (48 participants). Compensation was provided in the form of USD 0.70 (GBP 0.53) for completing Part I with an equivalent amount as a bonus payment for returning and completing Part II.

Table 9.2 Demographic characteristics (frequency and %) of participants, by group and overall

Characteristics		Groups of participant				Overall (N=220)
		Group 1 (N=57)	Group 2 (N=56)	Group 3 (N=59)	Group 4 (N=48)	
Gender	<i>Female</i>	27 (47.37)	24 (42.86)	27 (45.76)	19 (39.58)	97 (41.09)
	<i>Male</i>	30 (52.63)	32 (57.14)	32 (54.24)	29 (60.42)	123 (55.91)
Language	<i>English</i>	36 (63.16)	31 (55.36)	40 (67.80)	31 (64.58)	138 (62.73)
	<i>Other</i>	21 (36.84)	25 (44.64)	19 (32.20)	17 (35.42)	82 (37.27)
Education	<i>School</i>	4 (7.02)	-	-	3 (6.25)	7 (3.18)
	<i>Diploma</i>	8 (14.04)	6 (10.71)	8 (13.56)	5 (10.42)	27 (12.27)
	<i>Bachelor's</i>	28 (49.12)	28 (50.00)	27 (45.76)	26 (54.17)	109 (49.55)
	<i>Master's</i>	16 (28.07)	22 (39.29)	24 (40.68)	12 (25.00)	74 (33.64)
	<i>Doctoral</i>	1 (1.75)	-	-	2 (4.17)	3 (1.36)
Major/ Career	<i>Computing</i>	18 (31.58)	26 (46.43)	24 (40.68)	21 (43.75)	89 (40.45)
	<i>Non-computing</i>	39 (68.42)	30 (53.57)	35 (59.32)	27 (56.25)	131 (59.55)

Table 9.2 summarises the demographic characteristics of the participants both by group and overall. In terms of the entire sample, 97 (44.09%) participants were female and 123 (55.91%) were male. The participants ranged in age from 19 to 69 years, with a mean age of 36.86 years (standard deviation = 10.77). A majority of participants (138, 62.73%) were native speakers of English; the others had been speaking English for a mean of 25.16 years (standard deviation = 11.35). Almost half of the participants (109, 49.55%) had a bachelor's degree. The remaining participants' education levels ranged from a postgraduate degree (77, 35%) to a school qualification (7, 3.18%). In general, most participants had a non-computing major/career background (131, 59.55%). Moreover, a majority of participants spent on average more than six hours a day online using computers. As Table 9.2 demonstrates, the percentage distributions are almost the same across the four groups as they are for the entire sample. The exception was native language and major/career background, which featured differences below 10% between both English and non-English native speakers and computing and non-computing fields in *Group 2*.

9.2.3 Materials

Similar to Studies 4 and 6, two web-based applications were developed: a password creation application for Part I and a password recall application for Part II. The difference between this study's applications and those of the previous studies relates to the design of password creation pages that presented the PCS. This section discusses the design and structure of the current study's applications.

9.2.3.1 Password Creation Application

Figure 9.1 illustrates the structure of the password creation application designed for Part I. The application overall structure was similar to the structure discussed in studies 4 and 6 (see Section 6.2.3.1 in Chapter 6 and Section 8.2.3.1 in Chapter 8, respectively). The application started with the homepage and scenario pages. The online bank account scenario was used in this study, including the idea that participants had to imagine the need to create a new password for a compromised account.

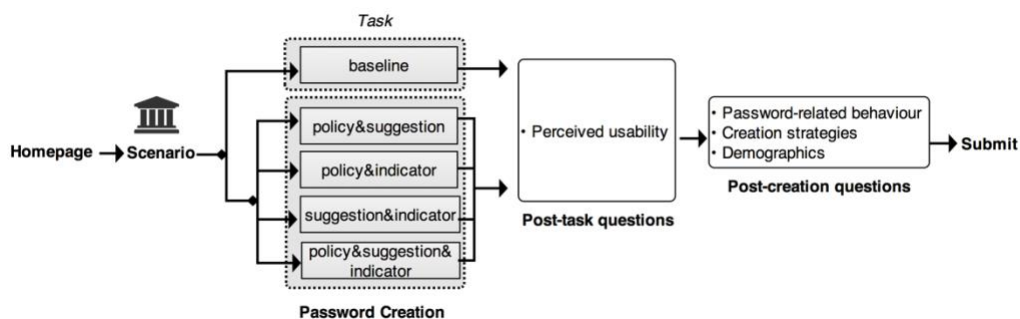


Figure 9.1 Structure of the password creation application used in the creation part

As Figure 9.1 demonstrates, five password creation pages were developed: *baseline*, *policy&suggestion*, *policy&indicator*, *suggestion&indicator*, and *policy&suggestion&indicator*. Once participants completed the password creation task (i.e. creating a new password and confirm it), an acknowledgement message popped up. For each password creation task, participants then completed (two in this study) a post-task page which appeared to collect information about the PCS. Thereafter a post-creation page appeared to ask participants to provide information about themselves. The following sections discuss the design and

contents of the password creation page (Section 9.2.3.1.1), the post-task page (Section 9.2.3.1.2), and the post-creation page (Section 9.2.3.1.3).

9.2.3.1.1 Password Creation Page

All password creation pages contained two fields: one to enter the password and one to confirm it. This was the design for the *baseline* page. Additional elements were included in the password creation page. These elements varied depending on the type of combination condition: the *policy&suggestion* page featured password policy and creation suggestion statements, the *policy&indicator* page included a password policy statement and password strength indicator, the *suggestion&indicator* page used a creation suggestion statement and password strength indicator, and the *policy&suggestion&indicator* page included all three supporting features.

The *baseline* page was always the first password creation page to be presented in the application, as depicted in Figure 9.1. This page did not include any supporting features to help the participant create a new password.

The four remaining pages used the same policy and creation suggestion statements as in Study 6 (see Section 8.2.3.1.1, Chapter 8), along with the algorithm for the password strength indicator. The following was used for the policy statement: ‘*The password needs to have at least twelve characters and at least three of the four character classes: uppercase letters, lowercase letters, numbers, or symbols.*’ The creation suggestion statement read as follows: ‘*You can improve your password by having a combination of numbers, letters and symbols like ! @ # { ; ~.*’ Finally, the password strength indicator used a green-amber-red traffic light metaphor with textual and graphical representation. Figures 9.2 to 9.4 illustrate how the PCS behaved in the four password creation pages across the three timings of presentation. types of combination conditions. The design and components of the four password creation pages are summarised, as follows:

- *policy&suggestion* page (see Figure 9.2): participants were provided with both the password policy and creation suggestion statements. The policy statement was presented before the participants started to create a password and remained

in the page until the point of password submission. In contrast, the suggestion statement was presented as the participants started to enter a password but disappeared once they moved on to the confirm password field.

- *policy&indicator* page (see Figure 9.3): participants received both the password policy statement and strength indicator. The policy statement was presented in the same way as on the *policy&suggestion* page; however, this time it was combined with a password strength indicator. The strength indicator was presented using a green-amber-red traffic light metaphor with textual and graphical representation.
- *suggestion&indicator* page (see Figure 9.4): participants were given both the suggestion statement and strength indicator. The suggestion statement was presented in the same way as on the *policy&suggestion* page, whereas the strength indicator was presented in the same way as on the *policy&indicator* page.
- *policy&suggestion&indicator* page (see Figure 9.5): all three features were combined and presented to participants in the same way as on the previous pages.

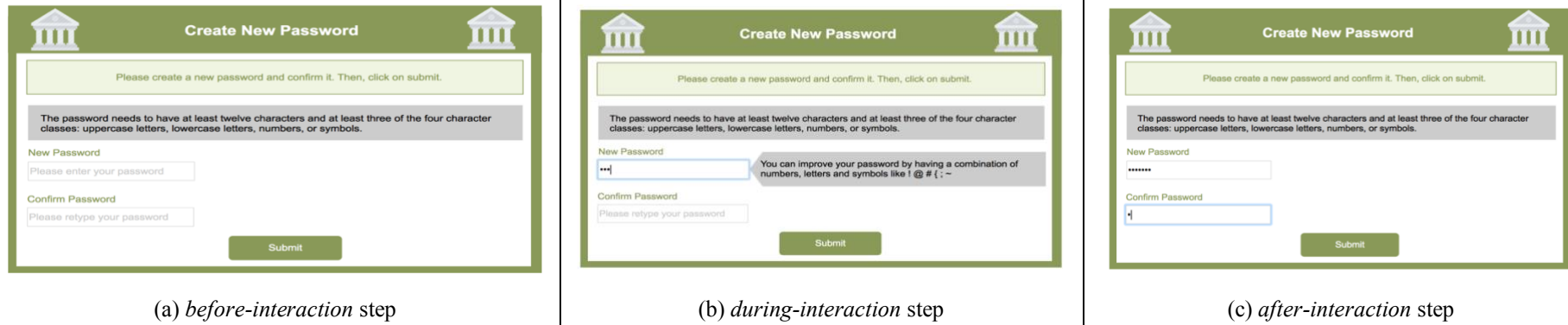


Figure 9.2 Screenshots of the design provided on the *policy&suggestion* password creation page across the three timings of presentation

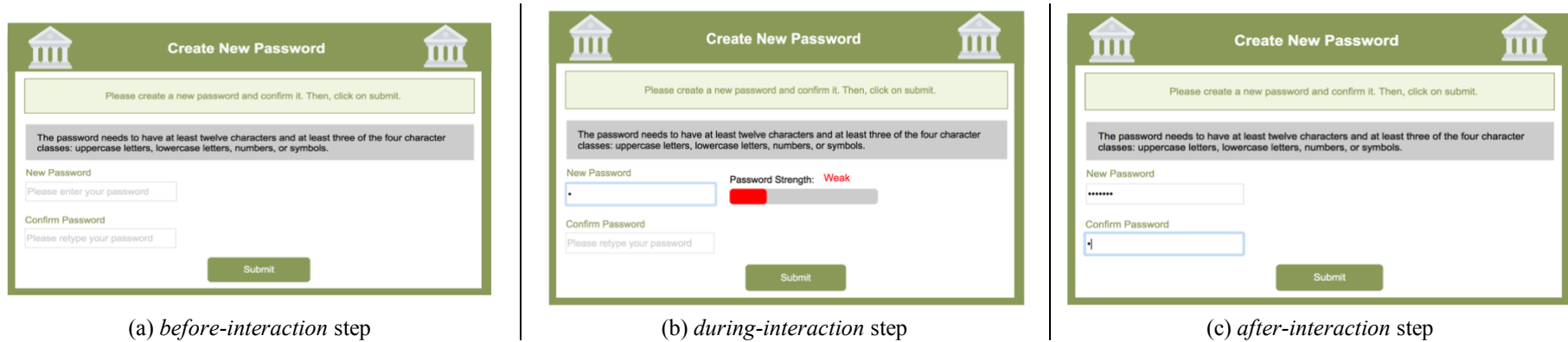
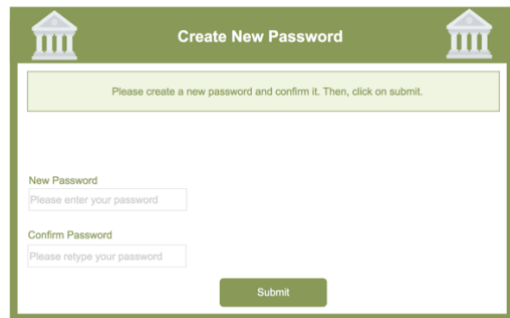
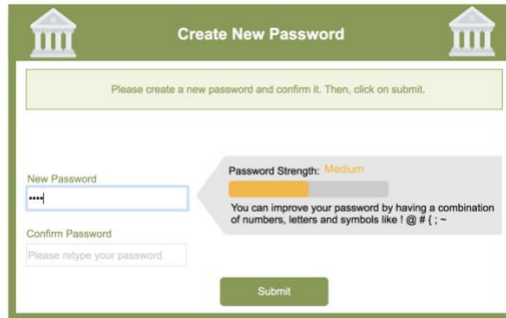


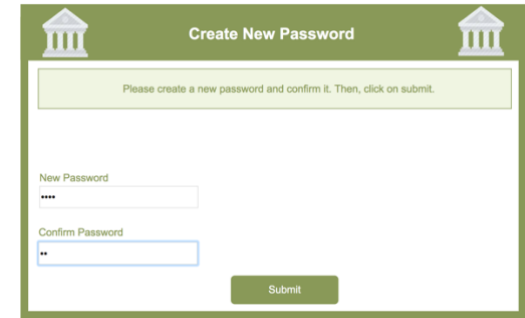
Figure 9.3 Screenshots of the design provided on the *policy&indicator* password creation page across the three timings of presentation



(a) before-interaction step

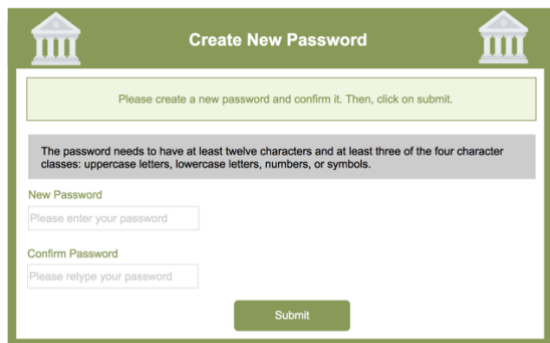


(b) during-interaction step

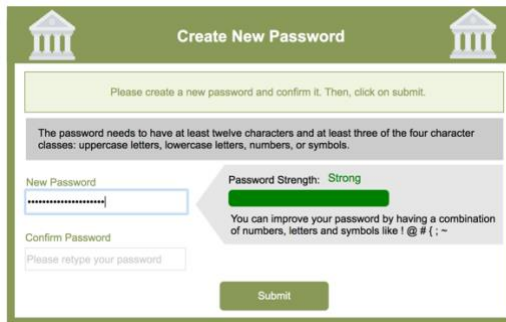


(c) after-interaction step

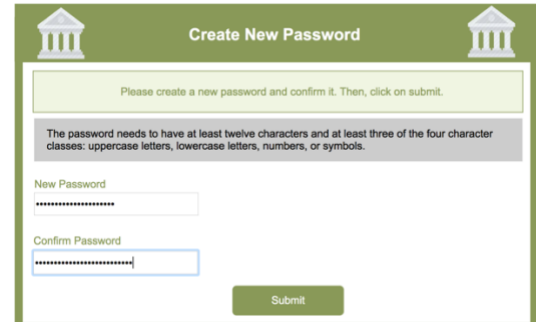
Figure 9.4 Screenshots of the design provided on the *suggestion&indicator* password creation page across the three timings of presentation



(a) before-interaction step



(b) during-interaction step



(c) after-interaction step

Figure 9.5 Screenshots of the design provided on the *policy&suggestion&indicator* password creation page across the three timings of presentation

9.2.3.1.2 Post-Task Questions Page

This page contained questions about user satisfaction in relation to the PCS that were similar to the questions provided in Study 6 (see Section 8.2.3.1.2, Chapter 8). They asked about the ease of use, annoyingness, helpfulness, clarity, and amount of detail, in addition to the participants' confidence in using the PCS. These variables were measured using a five-point Likert item ranging from 1 to 5, with higher scores being better. An optional open-ended question also gave participants a chance to explain their ratings of the PCS. This page was provided at the end of the password creation tasks.

9.2.3.1.3 Post-Creation Questions Page

This page contained questions about participants' password-related behaviours, the password creation strategies they used to create their passwords in this study, and participants' demographic characteristics. It was similar to the one used in Study 6 (see Section 8.2.3.1.3, Chapter 8) and provided once the participant completed the password creation application.

9.2.3.2 Password Recall Application

A recall page was developed for every password creation page. Each recall page had the following overall structure: task instructions, a screenshot of the PCS used to create the password, a password entry field, and a question about the participant's confidence in recalling the correct password. At the end of Part II, a post-recall questions page was presented to ask participants about the methods they used to remember their created passwords and their password management strategies. This application was similar to the those developed in Studies 4 and 6 (see Section 6.2.3.2 in Chapter 6 and Section 8.2.3.2 in Chapter 8, respectively).

9.2.4 Pilot of the Study Procedure

Three PhD students from the Computer Science Department took part in a pilot study to test the study procedure and the provided instructions. Both were perceived as being clear to follow, and no issues were reported about the study procedure. The data from the pilot were not included in the data analysis.

9.2.5 Procedure

The same procedure as in Studies 4 and 6 (see Section 6.2.5 in Chapter 6 and Section 8.2.5 in Chapter 8, respectively) was followed in the present study, since all three studies used the MTurk platform for participant recruitment. Each participant was assigned randomly to *Group 1*, *Group 2*, *Group 3*, or *Group 4*. Participants were directed to the password creation application through MTurk; three days later, they were invited via email to return and recall their passwords.

9.2.6 Data Analysis

Kolmogorov-Smirnov and Shapiro-Wilk tests were used to test for normality on all dependent measures. Most of these measures were significantly non-normal ($p < 0.05$) for both tests, with the exception of the creation time and number of lowercase letters. As such, non-parametric statistics were used throughout the analysis. During the data preparation process, the author identified and adjusted for outliers for the following dependent measures: creation time, keystrokes, password length, number of digits, number of uppercase letters, number of lowercase letters, number of symbols, password strength score, and recall time. The same method mentioned in Study 4 (see Section 6.2.6, Chapter 6) was used to adjust the outliers in the present study.

The data was examined using between-participants and within-participant analyses. The between-participants analysis, which used Kruskal-Wallis tests (H statistic), was employed to compare participant performance between the four types of combinations conditions (i.e. *policy&suggestion*, *policy&indicator*, *suggestion&indicator*, and *policy&suggestion&indicator*). The within-participants analysis was utilised to

compare participant performance between the *baseline* and *combination* conditions; Wilcoxon signed-ranks tests (*Z* statistic) were used to this end.

Furthermore, when the dependent measures were of a frequency type, chi-square (χ^2 statistics) tests were used to measure the association among categories (i.e. number of character classes, policy compliance, suggestion compliance, password guessability, and accuracy).

9.3 Results

The results of the current study are divided into two sections: the password creation results from Part I and the password recall results from Part II. Thereafter a comparison between the individual (Study 6, Chapter 8) and combined effects (Study 7, present chapter) is presented for each supporting: password policy, creation suggestions, and strength indicators¹⁰.

9.3.1 Password Creation

This section presents results regarding the usability of the four types of combination conditions. It also presents findings related to the strength of the created passwords with these conditions.

¹⁰ If half of the items in a dependent measure (e.g. user satisfaction) showed a significant effect that was not considered enough evidence to conclude that there was an effect for that dependent measure.

9.3.1.1 Usability of PCS

9.3.1.1.1 Efficiency Measures

Table 9.3 summarises the results for the two efficiency measures, whereas Table 9.4 presents the results of the pairwise comparison undertaken for the two efficiency measures.

Table 9.3 Mean (median) creation time and keystrokes measures for the four types of combination conditions

	Types of combination conditions				<i>p</i> <i>value</i>
	<i>policy& suggestion</i> (Group 1)	<i>policy& indicator</i> (Group 2)	<i>suggestion& indicator</i> (Group 3)	<i>policy&suggestio n&indicator</i> (Group 4)	
Creation time	28.74 (26.00)	38.05 (35.50)	29.72 (28.00)	42.35 (36.50)	.016
Keystrokes	17.30 (15.00)	21.82 (19.00)	17.61 (15.00)	21.36 (17.00)	.019

Creation time. There was a significant difference in the creation time between the types of combination conditions ($H(3) = 10.33$, $p = .016$). Participants in the *policy&suggestion* and *suggestion&indicator* condition spent significantly less time creating passwords than those in the other two conditions. The pairwise comparison confirmed this and revealed no significant difference in the creation time between *policy&suggestion* and *suggestion&indicator* (see Table 9.4). In addition, there was a significant difference in the creation time between the *combination* and *baseline* conditions ($Z = -5.35$, $p < .001$): participants spent significantly more time creating passwords with the *combination* condition ($M = 34.34$, $Mdn = 31.00$) than with the *baseline* condition ($M = 25.79$, $Mdn = 21.00$).

Keystrokes. There was a significant difference in the number of keystrokes that participants used to create a password between the types of combination conditions ($H(3) = 9.99$, $p = .019$). Participants in the *policy&indicator* condition performed significantly more keystrokes than those in the other three conditions. The pairwise comparison revealed a significant difference between both the *policy&indicator* and *policy&suggestion* conditions and the *policy&indicator* and *suggestion&indicator* conditions (see Table 9.4). There was also a significant difference in the number of

keystrokes between the *combination* and *baseline* conditions ($Z = -4.99$, $p < .001$): participants performed significantly more keystrokes in the *combination* condition ($M = 19.42$, $Mdn = 16.00$) than in the *baseline* condition ($M = 15.29$, $Mdn = 13.50$).

Table 9.4 Pairwise comparisons of creation time and keystrokes measures across the four types of combination conditions

		<i>policy&sug gestion</i>	<i>policy& indicator</i>	<i>suggestion &indicator</i>	<i>policy&suggestio n&indicator</i>
Creation time	<i>policy&suggestion</i>	-	-26.87*	-2.58	-30.90*
	<i>policy&indicator</i>		-	24.29*	-4.03
	<i>suggestion&indicator</i>			-	-28.32*
	<i>policy&suggestion&indicator</i>				-
Keystrokes	<i>policy&suggestion</i>	-	-29.59*	0.05	-23.51
	<i>policy&indicator</i>		-	29.63*	6.07
	<i>suggestion&indicator</i>			-	-23.56
	<i>policy&suggestion&indicator</i>				-

Note. * denotes a significant pairwise comparison result, $p < .05$.

9.3.1.1.2 User Satisfaction Measures

Table 9.5 summarises the results for the six user satisfaction measures (i.e. ease of use, annoyingness, helpfulness, clarity, amount of detail, and confidence).

Table 9.5 Mean (median) ratings of user satisfaction measures across the four types of combination conditions

	Types of combination conditions				<i>p value</i>
	<i>policy& suggestion (Group 1)</i>	<i>policy& indicator (Group 2)</i>	<i>suggestion& indicator (Group 3)</i>	<i>policy&suggestion &indicator (Group 4)</i>	
Ease of use	3.58 (4.00)	3.52 (4.00)	3.66 (4.00)	3.75 (4.00)	n.s.
Annoyingness	3.91 (4.00)	3.71 (4.00)	3.68 (4.00)	4.02 (4.00)	n.s.
Helpfulness	3.65 (4.00)	4.00 (4.00)	3.88 (4.00)	3.81 (4.00)	n.s.
Clarity	3.96 (4.00)	4.09 (4.00)	4.02 (4.00)	3.96 (4.00)	n.s.
Amount of detail	2.54 (3.00)	2.46 (3.00)	2.44 (3.00)	2.73 (3.00)	n.s.
Confidence	4.05 (4.00)	4.04 (4.00)	3.75 (4.00)	3.88 (4.00)	n.s.

Ease of use. There was no significant difference in the ratings for ease of use between the four types of combination conditions ($H(3) = 1.17$, $p = .760$). However, there was

a significant difference in the ease of use ratings ($Z = -4.60$, $p < .001$) between the *combination* and *baseline* conditions. Unexpectedly, participants perceived the presentation of a combined supporting feature ($M = 3.62$, $Mdn = 4.00$) as making password creation more difficult compared to the baseline ($M = 3.99$, $Mdn = 4.00$).

Annoyingness. There was no significant difference in the ratings for annoyingness between the four types of combination conditions ($H(3) = 1.78$, $p = .618$). However, there was a significant difference in the annoyingness ratings ($Z = -5.28$, $p < .001$) between the *combination* and *baseline* conditions: participants perceived the provision of a combined supporting feature as more annoying ($M = 3.82$, $Mdn = 4.00$) than the baseline ($M = 4.23$, $Mdn = 5.00$).

Helpfulness. There was no significant difference in the ratings for helpfulness between the four types of combination conditions ($H(3) = 2.53$, $p = .470$). However, there was a significant difference in the helpfulness ratings ($Z = -5.46$, $p < .001$) between the *combination* and *baseline* conditions: participants perceived the provision of a combination of supporting features ($M = 3.84$, $Mdn = 4.00$) when creating password as more helpful than their non-provision ($M = 3.35$, $Mdn = 4.00$).

Clarity. There was no significant difference in the ratings for clarity between the four types of combination conditions ($H(3) = 0.83$, $p = .842$). There was also no significant difference in the clarity ratings ($Z = -0.67$, $p = .512$) between the *combination* and *baseline* conditions.

Amount of detail. There was no significant difference in the ratings for the amount of detail between the four types of combination conditions ($H(3) = 4.88$, $p = .181$). However, there were significant differences in the amount of detail ratings ($Z = -3.19$, $p = .001$) between the *combination* and *baseline* conditions. Participants rated the amount of detail presented when a combined supporting feature was provided ($M = 2.54$, $Mdn = 3.00$) higher than the *baseline* condition ($M = 2.35$, $Mdn = 3.00$).

Confidence. There was no significant difference in the ratings for participants' confidence in creating passwords between the four types of combination conditions ($H(3) = 4.01, p = .260$). However, there were significant differences in participants' confidence ratings between the *combination* and *baseline* conditions ($Z = -3.06, p = .002$). Participants felt more confident creating passwords when a combined supporting feature was provided ($M = 3.93, Mdn = 4.00$) than when no such combined feature was presented ($M = 3.73, Mdn = 4.00$).

In summary, the PCS usability differed significantly among the four types of combination conditions in terms of only PCS efficiency. Participants used significantly less time and effort to create passwords when the PCS provided the suggestion statement in combination with either the policy statement or strength indicator (i.e. *policy&suggestion* and *suggestion&indicator*). Furthermore, PCS usability did differ significantly between the *combined* and *baseline* conditions in terms of PCS efficiency and user satisfaction (with the exception of clarity). Providing a combination of supporting features improved the level of user satisfaction (apart from ease of use), but it made the password creation process less efficient and more difficult to use.

9.3.1.2 Strength of Password

9.3.1.2.1 Password Characteristics

Table 9.6 summarises the results for the password characteristic measures (i.e. password length, number of digits, number of uppercase letters, number of lowercase letters, and number of symbols); Table 9.7 presents the results of the pairwise comparison undertaken these measures.

Password length. There was a significant difference in the length of passwords between the four types of combination conditions ($H(3) = 23.57, p < .001$). Passwords created in the *policy&suggestion&indicator* condition were longer than those created in the other three conditions. The pairwise comparison revealed a significant

difference in length between passwords created in the *policy&suggestion&indicator* and *policy&indicator* conditions and those formulated in the *policy&suggestion* and *suggestion&indicator* conditions (see Table 9.7). However, no significant difference in password length was found between the *policy&suggestion&indicator* and *policy&indicator* conditions. Furthermore, there was a significant difference in password length between the *combination* and *baseline* conditions ($Z = -8.56$, $p < .001$): participants created longer passwords when they were provided with a combined supporting feature ($M = 12.23$, $Mdn = 12.00$) in comparison to the baseline ($M = 10.35$, $Mdn = 10.00$).

Table 9.6 Mean (median) ratings of password characteristic measures for the four types of combination conditions

	Types of combination conditions				<i>p</i> value
	<i>policy&suggestion</i> (Group 1)	<i>policy&indicator</i> (Group 2)	<i>suggestion&indicator</i> (Group 3)	<i>policy&suggestion&indicator</i> (Group 4)	
Password length	11.79(12.00)	12.82(13.00)	11.42 (11.00)	13.06 (13.00)	.000
Number of digits	3.37 (4.00)	2.63 (3.00)	2.63 (3.00)	3.08 (4.00)	.029
Number of uppercase	1.25 (1.00)	1.58 (1.00)	0.74 (1.00)	1.18 (1.00)	.006
Number of lowercase	5.67 (6.00)	7.16 (7.00)	6.38 (6.00)	6.88 (6.00)	.037
Number of symbols	1.19 (1.00)	1.09 (1.00)	1.02 (1.00)	0.80 (1.00)	n.s.

Number of digits. There was a significant difference in the number of digits used in passwords between the four types of combination conditions ($H(3) = 9.02$, $p = .029$). Passwords created in the *policy&suggestion* condition had significantly more digits than those formed in the *policy&indicator* and *suggestion&indicator* conditions, as confirmed by the pairwise comparison (see Table 9.7). There was also a significant difference in the number of digits used in the *combination* and *baseline* conditions ($Z = -2.99$, $p = .003$): passwords created with a combined supporting feature ($M = 2.92$, $Mdn = 3.00$) had more digits than those created without a supporting feature ($M = 2.59$, $Mdn = 2.00$).

Number of uppercase letters. There was a significant difference in the number of uppercase letters used in passwords between the four types of combination conditions

($H(3) = 12.42, p = .006$). Passwords created in the *suggestion&indicator* condition had fewer uppercase letters than those created in the other three conditions. This was confirmed by the pairwise comparison, but no significant difference in the number of uppercase letters was found among the other three conditions (see Table 9.7). Furthermore, there was a significant difference in the number of uppercase letters used in the *combination* and *baseline* conditions ($Z = -6.11, p < .001$): passwords created with a combined supporting feature ($M = 1.18, Mdn = 1.00$) had more uppercase letters than which such a feature was not provided ($M = 0.72, Mdn = 1.00$).

Table 9.7 Pairwise comparison of password characteristic measures between the four types of combination conditions

		<i>policy&suggestion</i>	<i>policy&indicator</i>	<i>suggestion&indicator</i>	<i>policy&suggestion&indicator</i>
Password length	<i>policy&suggestion</i>	-	-31.63*	9.90	-39.36*
	<i>policy&indicator</i>		-	41.52*	-7.73
	<i>suggestion&indicator</i>			-	-49.26*
	<i>policy&suggestion&indicator</i>				-
Number of digits	<i>policy&suggestion</i>	-	28.32*	29.24*	9.46
	<i>policy&indicator</i>		-	0.92	-18.86
	<i>suggestion&indicator</i>			-	-19.78
	<i>policy&suggestion&indicator</i>				-
Number of uppercase	<i>policy&suggestion</i>	-	-13.50	25.13*	-3.08
	<i>policy&indicator</i>		-	38.63*	10.43
	<i>suggestion&indicator</i>			-	-28.21*
	<i>policy&suggestion&indicator</i>				-
Number of lowercase	<i>policy&suggestion</i>	-	-33.14*	-17.15	-25.37*
	<i>policy&indicator</i>		-	15.99	7.77
	<i>suggestion&indicator</i>			-	-8.23
	<i>policy&suggestion&indicator</i>				-

Note. * denotes a significant pairwise comparison result, $p < .05$.

Number of lowercase letters. There was a significant difference in the number of lowercase letters used in passwords between the four types of combination conditions ($H(3) = 8.48, p = .037$). Passwords created in the *policy&suggestion* condition had fewer lowercase letters than those created in the *policy&indicator* and *policy&suggestion&indicator* conditions, as confirmed by the pairwise comparison

(see Table 9.7). Furthermore, there was a significant difference in the number of lowercase letters included in passwords in the *combination* and *baseline* conditions ($Z = -3.28$, $p = .001$): passwords created with a combined supporting feature ($M = 6.50$, $Mdn = 6.00$) had more lowercase letters than those created when such a feature was not provided ($M = 5.68$, $Mdn = 5.00$).

Number of symbols. There was no significant difference in the number of symbols used in passwords between the four types of combination conditions ($H(3) = 4.62$, $p = .202$). However, there was a significant difference in the number of symbols used in passwords in the *combination* and *baseline* conditions ($Z = -7.58$, $p < .001$): passwords created with a combined supporting feature ($M = 1.03$, $Mdn = 1.00$) had more symbols than those formulated in the absence of such a feature ($M = 0.54$, $Mdn = 0.00$).

Number of password character classes. Counting the number of different character classes that occurred in passwords, there were significant differences in the number of classes used in passwords in all four types of combination conditions: *policy&suggestion* (Group 1: $\chi^2(2) = 14.00$, $p = .001$), *policy&indicator* (Group 2: $\chi^2(3) = 22.71$, $p < .001$), *suggestion&indicator* (Group 3: $\chi^2(2) = 24.86$, $p < .001$), and *policy&suggestion&indicator* (Group 4: $\chi^2(2) = 11.38$, $p = .003$). The distribution of the password character classes across the types of combination conditions is presented in Figure 9.6. The *policy&suggestion* condition had the highest percentage of passwords using all four character classes (54.39%), followed by *policy&suggestion&indicator* (45.73%), *policy&indicator* (42.86%), and *suggestion&indicator* (40.68%).

Furthermore, passwords created in the *combination* condition had significantly more character classes than those formulated in the *baseline* condition ($Z = -8.08$, $p < .001$). In the *combination* condition, almost half of all passwords included four character classes (45.91%); in the *baseline* condition, most passwords included three (25.45%) or four (30.00%) character classes.

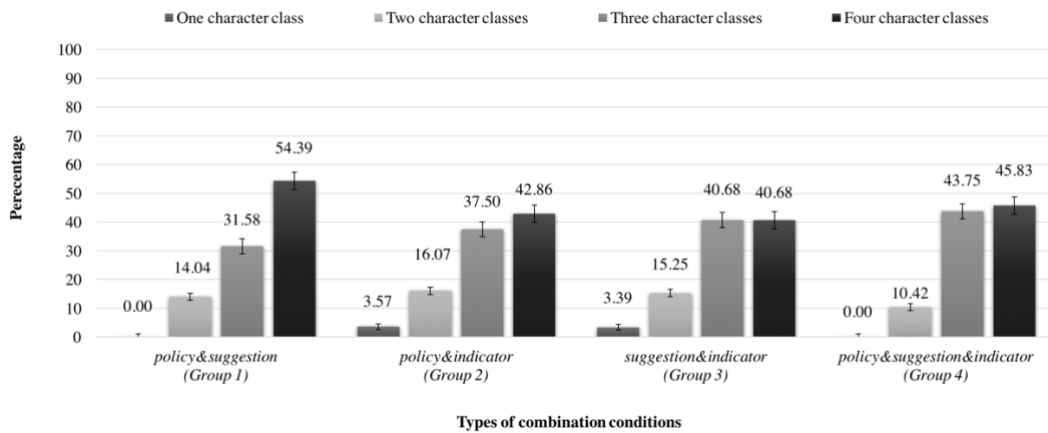


Figure 9.6 Percentage of password character classes across the four types of combination conditions **Policy compliance**. Only three of the four types of combination conditions were analysed for this measure: *policy&suggestion* (Group 1), *policy&indicator* (Group 2), and *policy&suggestion&indicator* (Group 4); the *suggestion&indicator* condition (Group 3) was excluded as it did not provide a policy statement. The results revealed that there was a significant difference in the number of passwords that followed the given policy in only one type of combination condition: *policy&suggestion&indicator* (Group 4: $\chi^2(1) = 8.33, p = .004$). No such difference existed in relation to the other two conditions analysed: *policy&suggestion* (Group 1: $\chi^2(1) = 0.86, p = .354$) and *policy&indicator* (Group 2: $\chi^2(1) = 2.57, p = .109$).

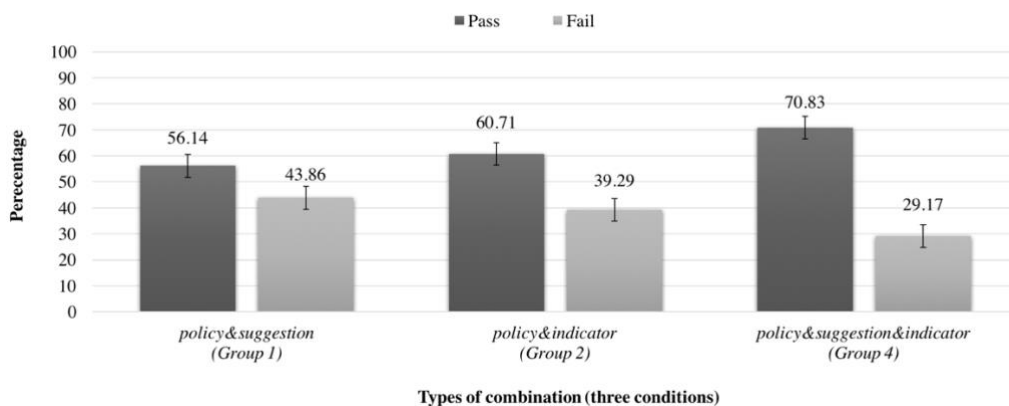


Figure 9.7 Percentage of policy compliance across three types of combination conditions

The distribution of policy compliance across the three types of combination conditions is shown in Figure 9.7. The *policy&suggestion&indicator* condition had the highest percentage of passwords that followed the policy (70.83%).

Suggestion compliance. Only three of the four types of combination conditions were analysed for this measure: *policy&suggestion* (Group 1), *suggestion&indicator* (Group 3), and *policy&suggestion&indicator* (Group 4); the *policy&indicator* condition (Group 2) was excluded as it did not include a suggestion statement. The results indicated that there was a significant difference in the number of passwords that followed the given suggestion in two types of combination conditions: *policy&suggestion* (Group 1: $\chi^2(1) = 7.74$, $p = .005$) and *suggestion&indicator* (Group 3: $\chi^2(1) = 8.97$, $p = .003$); a significant difference was not found in relation to the *policy&suggestion&indicator* condition (Group 4: $\chi^2(1) = 2.08$, $p = .149$) condition. The distribution of suggestion compliance across the three types of combination conditions is presented in Figure 9.8. The *suggestion&indicator* condition had the highest percentage of passwords that followed the suggestion (69.49%), followed by *policy&suggestion* (68.42%).

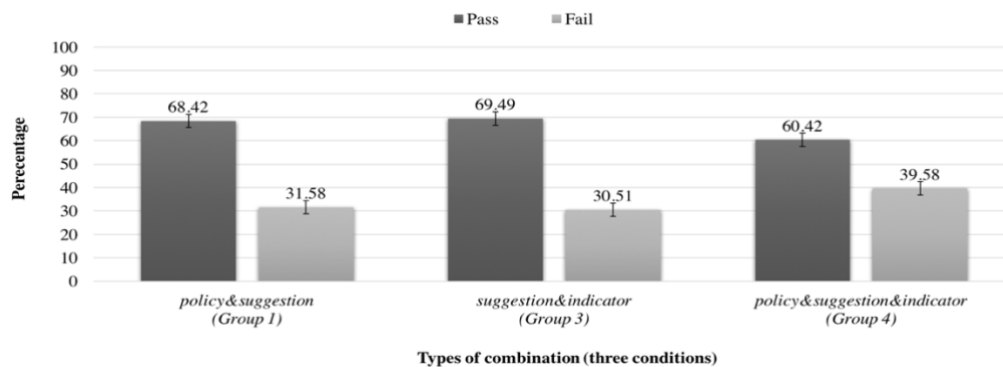


Figure 9.8 Percentage of suggestion compliance across three types of combination conditions

Symbols provision. The same three types of combination conditions analysed in relation to suggestion compliance were analysed for this measure: *policy&suggestion* (Group 1), *suggestion&indicator* (Group 3), and *policy&suggestion&indicator*

(Group 4). The results indicated that no significant difference in the number of passwords that included the provided symbols in all three types of combination conditions: *policy&suggestion* (Group 1: $\chi^2(1) = 1.42$, $p = .233$), *suggestion&indicator* (Group 3: $\chi^2(1) = 2.86$, $p = .091$), and *policy&suggestion&indicator* (Group 4: $\chi^2(1) = 1.33$, $p = .248$). The percentage of passwords that included at least one of the symbols across the three types of combination conditions is presented in Figure 9.9.

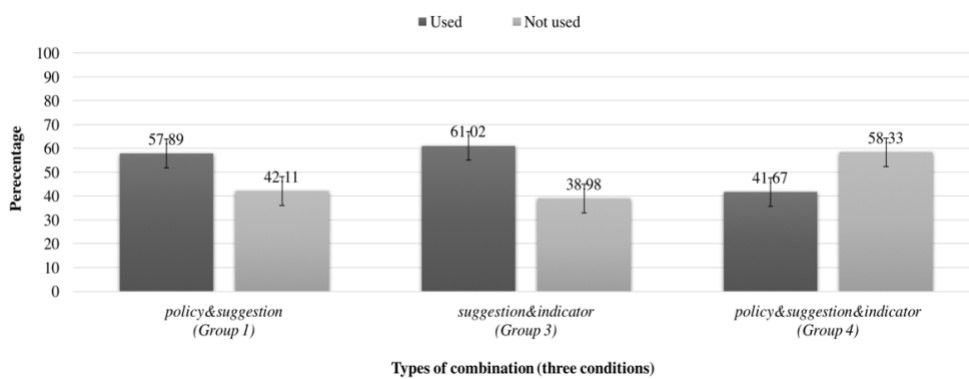


Figure 9.9 Percentage of passwords that included the given symbols across three types combination conditions

Password strength score. Only three types of combination conditions were analysed for this measure: *policy&indicator* (Group 2), *suggestion&indicator* (Group 3), and *policy&suggestion&indicator* (Group 4); the *policy&suggestion* condition (Group 1) was excluded as it did not provide the password strength indicator. The results revealed that a significant difference in password strength scores was found between the three types of combination conditions ($H(2) = 15.16$, $p = .001$). Passwords created with the *suggestion&indicator* condition ($M = 72.28$, $Mdn = 72.27$) were weaker than those formed using the other two conditions: *policy&indicator* ($M = 78.47$, $Mdn = 78.84$) and *policy&suggestion&indicator* ($M = 80.71$, $Mdn = 80.78$). The pairwise comparison confirmed this difference. It also showed no significance in the password strength scores between *policy&indicator* and *policy&suggestion&indicator*, as indicated in Table 9.8. The distribution of password strength levels across the different conditions is shown in Figure 9.10.

Table 9.8 Pairwise comparison of password strength scores across three types of combination conditions

		<i>policy&indicator</i>	<i>suggestion&indicator</i>	<i>policy&suggestion&indicator</i>
Password strength score	<i>policy&indicator</i>	-	25.60*	-7.82
	<i>suggestion&indicator</i>		-	-33.42*
	<i>policy&suggestion&indicator</i>			-

Note. * denotes a significant pairwise comparison result, $p < .05$.

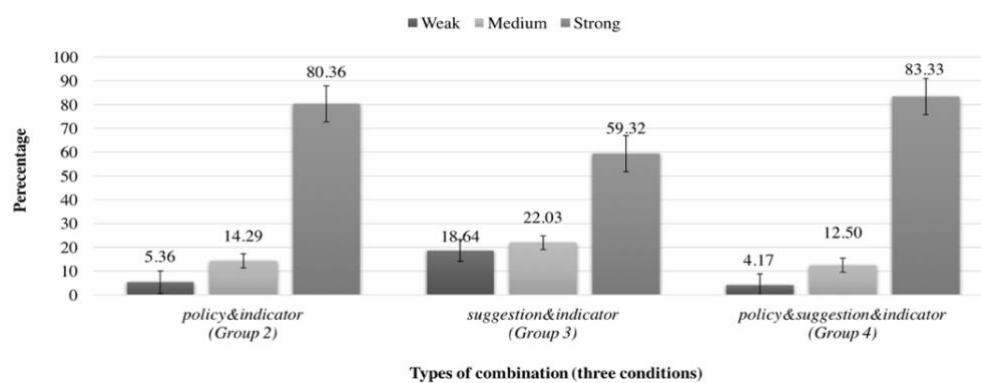


Figure 9.10 Percentage of password strength levels across three types of combination conditions

9.3.1.2.2 Password Guessability

Out of the 220 passwords created by participants, 159 complied with the study's password policy. The ability to guess these 159 passwords using five cracking approaches, (based on Ur et al., 2015), was also used to measure password strength. Overall, more than half of the compliant passwords (100, 62.89%) were not guessable and just over a quarter (59, 26.82%) were guessable using at least one cracking approach. Figure 9.11 presents the password guessability percentages across the four types of combination conditions.

The results revealed that there was a significant difference in the number of guessable and non-guessable passwords in only two conditions: *policy&suggestion* (Group 1: $\chi^2(1) = 6.13$, $p = .013$) and *policy&indicator* (Group 2: $\chi^2(1) = 9.53$, $p = .002$). No significant difference was found in the number of guessable and non-guessable passwords in the *suggestion&indicator* (Group 3: $\chi^2(1) = 0.02$, $p = .896$) and

policy&suggestion&indicator (Group 4: $x^2(1) = 1.88$, $p = .170$) conditions. The *policy&indicator* condition had the highest percentage of passwords that were not guessable (76.47%), followed by the *policy&suggestion* condition (71.88%).

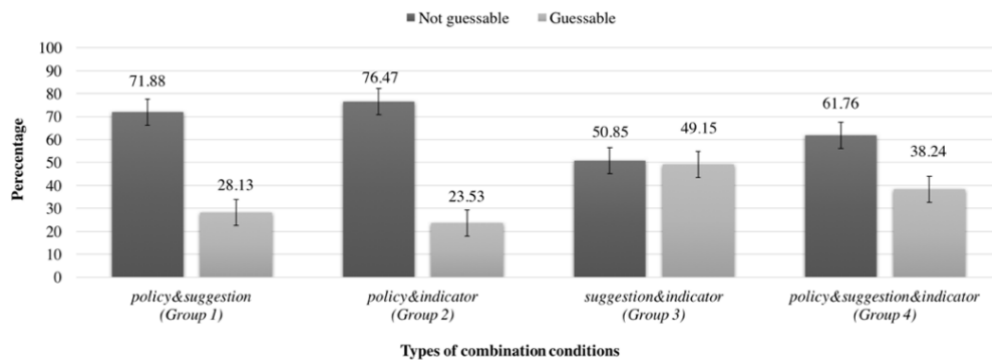


Figure 9.11 Percentages of password guessability across the four types of combination conditions

In summary, the strength of the created passwords did differ significantly among the four types of combination conditions in relation to password length and the number of digits, uppercase letters, and lowercase letters included. Passwords were significantly longer when they were created using either all supporting features or the policy statement in combination with the strength indicator. Passwords included a higher number of uppercase and lowercase letters when any combination condition was used except when the suggestion statement was combined with either the strength indicator or password policy. Moreover, they included a higher number of digits when they were created using either all supporting features or the policy statement combined with the creation suggestion. Most passwords were not guessable when the policy statement was combined with either the creation suggestion statement or the password strength indicator. Furthermore, the password strength did differ significantly between the *combined* and *baseline* conditions in terms the password characteristics. Providing a combination of supporting features improved the strength of the passwords.

9.3.2 Password Recall

Out of the 220 participants, only 147 (66.82%) completed the recall task within 24 hours of the invitations being sent: 37 in the *policy&suggestion* condition (Group 1),

44 in the *policy&indicator* condition (Group 2), 33 in the *suggestion&indicator* condition (Group 3), and 33 in the *policy&suggestion&indicator* condition (Group 4). Based on this sample, this section presents the results regarding PCS usability when users recall their passwords.

9.3.2.1.1 Efficiency Measures

Recall time. There was no significant difference in the recall times between the four types of combination conditions ($H(3) = 5.22, p = .157$); Figure 9.12 presents the related mean recall times. However, there was a significant difference in the recall times between the *combination* and *baseline* conditions ($Z = -8.40, p < .001$). The recall time for passwords created with a combined supporting feature ($M = 28.85, Mdn = 25.00$) was shorter than for those created in the baseline condition ($M = 45.21, Mdn = 45.00$).

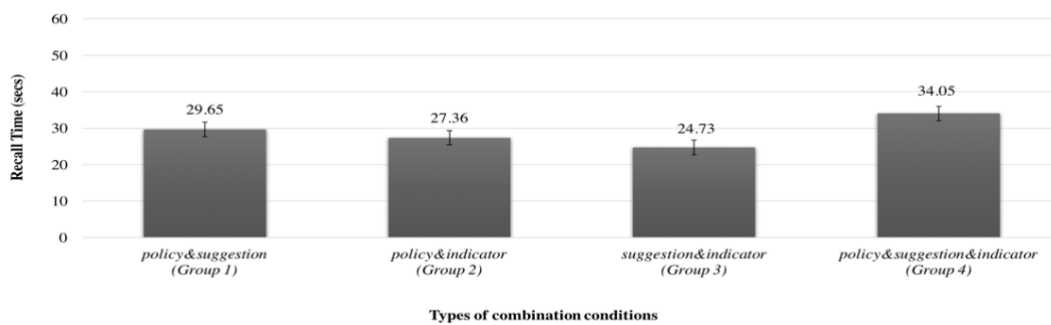


Figure 9.12 Mean recall times across the four types of combination conditions

9.3.2.1.2 Effectiveness Measures

Accuracy. Figure 9.13 shows the percentage of password recall accuracy for the four types of combination conditions. There was no significant difference in participants' accuracy in recalling passwords in all four conditions: *policy&suggestion* (Group 1: $\chi^2(1) = 0.24, p = .622$), *policy&indicator* (Group 2: $\chi^2(1) = 0.09, p = .763$), *suggestion&indicator* (Group 3: $\chi^2(1) = 0.03, p = .862$), and *policy&suggestion&indicator* (Group 4: $\chi^2(1) = 0.76, p = .384$). There was also no

significant difference in password recall accuracy between the *combination* and *baseline* conditions ($Z = -1.06$, $p = .289$).

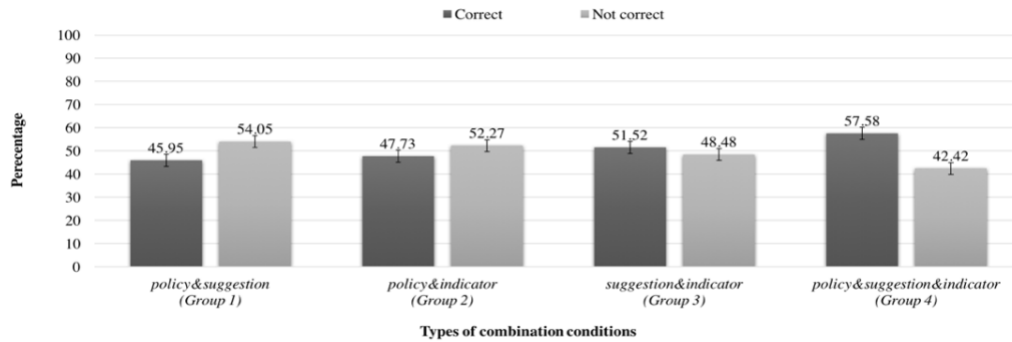


Figure 9.13 Percentage of password recall accuracy across the four types of combination conditions

9.3.2.1.3 User Satisfaction Measures

Confidence. There was no significant difference between the types of combination conditions in relation to how participants rated their confidence in recalling the correct password ($H(3) = 1.41$, $p = .703$); Figure 9.14 presents the related mean ratings. There was also no significant difference in participants' confidence ratings between the *combination* and *baseline* conditions ($Z = -1.19$, $p = .233$).

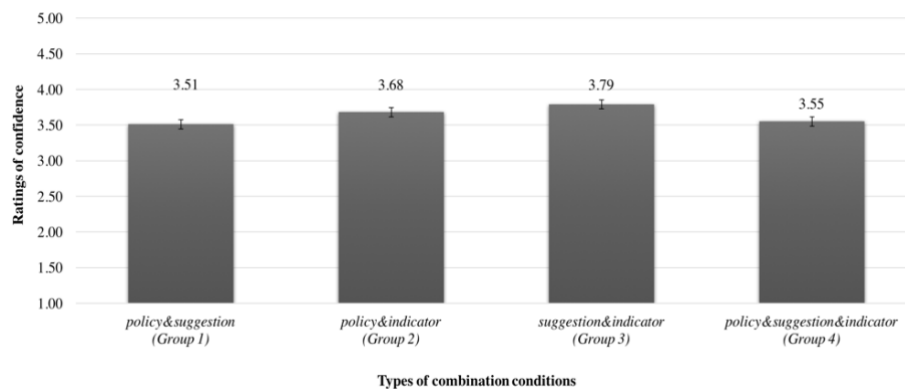


Figure 9.14 Mean ratings of participants' recall confidence across the four types of combination conditions

9.3.3 Comparison Between Individual and Combined Effects

The author assumed that providing more than one supporting feature during the password creation process would improve the level of PCS user satisfaction. However, this study's results indicated that the perceived usability of PCSs was not affected by different combinations of supporting features. The question then becomes to what extent providing an individual (single) supporting feature to a combined feature would affect the PCS usability and password strength. A comparison was thus conducted to examine the differences between individual effects (study 6 in chapter 8) and combined effects (study 7, present chapter).

Since the best presentation of each supporting feature found in Study 6 was used to design different combinations of features in the present study, the results from Study 6 were used to compare the individual and combined effects. To recall, the effective presentations used for the features in the present study were *policy-before-interaction*, *suggestion-during-interaction*, and *3colour-graphical&textual*; these three presentation conditions represented the individual effects (hereinafter referred to as *policy-only*, *suggestion-only*, and *indicator-only*, respectively). The comparison was performed for each supporting feature separately, as follows:

- **Policy:** four presentation conditions were examined: *policy-only* (68 participants), *policy&suggestion* (57 participants), *policy&indicator* (56 participants), and *policy&suggestion&indicator* (48 participants).
- **Creation suggestion:** four presentation conditions were investigated: *suggestion-only* (63 participants), *policy&suggestion* (57 participants), *suggestion&indicator* (59 participants), and *policy&suggestion&indicator* (48 participants).
- **Strength indicator:** four presentation conditions were considered: *indicator-only* (81 participants), *policy&indicator* (56 participants), *suggestion&indicator* (59 participants), and *policy&suggestion&indicator* (48 participants).

9.3.3.1 Password Creation

This section presents the results regarding PCS usability and password strength.

9.3.3.1.1 Usability of PCS

9.3.3.1.1.1 Efficiency Measures

Table 9.9 and Table 9.10 respectively summarise the between-participant results for creation time and keystrokes for all three supporting features. Table 9.11 illustrates the pairwise comparison results across the presentation conditions for the efficiency measures.

Table 9.9 Mean (median) creation time measures across different presentation conditions for policy, suggestion, and indicator

	<i>policy-only</i>	<i>suggestion-only</i>	<i>indicator-only</i>	<i>policy&suggestion</i>	<i>policy&indicator</i>	<i>suggestion&indicator</i>	<i>policy&suggestion&indicator</i>	<i>p value</i>
Policy	112.07 (35.00)	-	-	28.74 (26.00)	38.05 (35.50)	-	42.35 (36.50)	n.s.
Creation suggestion	-	85.28 (28.00)	-	28.74 (26.00)	-	29.72 (28.00)	42.35 (36.50)	n.s.
Strength indicator	-	-	71.44 (32.00)	-	38.05 (35.50)	29.72 (28.00)	42.35 (36.50)	n.s.

Creation time. There was no significant difference in the creation time between the presentation conditions in all three features (see Table 9.9): policy ($H(3) = 7.72$, $p = .052$), creation suggestion ($H(3) = 5.27$, $p = .153$), and strength indicator ($H(3) = 5.18$, $p = .134$).

Keystrokes. There was a significant difference in the number of keystrokes used to create a password between the presentation conditions in only one feature (see Table 9.10): strength indicator ($H(3) = 8.13$, $p = .043$). No significant difference was found in relation to the other two features: policy ($H(3) = 6.73$, $p = .081$) and creation suggestion ($H(3) = 4.70$, $p = .195$). Regarding the strength indicator feature, participants who created passwords in the *suggestion&indicator* condition used

significantly fewer keystrokes than those in the *indicator-only* and *policy&indicator* conditions. The pairwise comparison confirmed this difference (see Table 9.11).

Table 9.10 Mean (median) of keystrokes measures for policy, suggestion, and indicator across different presentation conditions

	<i>policy-only</i>	<i>suggestion-only</i>	<i>indicator-only</i>	<i>policy&suggestion</i>	<i>policy&indicator</i>	<i>suggestion&indicator</i>	<i>policy&suggestion&indicator</i>	<i>p value</i>
Policy	23.34 (17.00)	-	-	17.30 (15.00)	21.82 (19.00)	-	21.36 (17.00)	n.s.
Creation suggestion	-	20.54 (17.50)	-	17.30 (15.00)	-	17.61 (15.00)	21.36 (17.00)	n.s.
Strength indicator	-	-	23.42 (20.00)	-	21.82 (19.00)	17.61 (15.00)	21.36 (17.00)	.043

Table 9.11 Pairwise comparisons of keystrokes measures across different presentation conditions

Keystrokes	Strength indicator	<i>indicator-only</i>	<i>policy&indicator</i>	<i>suggestion&indicator</i>	<i>policy&suggestion&indicator</i>
		<i>indicator-only</i>	-	1.98	31.81*
	<i>policy&indicator</i>		-	29.83*	5.89
	<i>suggestion&indicator</i>			-	-1.76
	<i>policy&suggestion&indicator</i>				-

Note. * denotes a significant pairwise comparison result, $p < .05$.

9.3.3.1.1.2 User Satisfaction Measures

In this section, Tables 9.12, 9.14, 9.16, 9.18, 9.19, and 9.21 summarise the between-participants results for all three supporting features. Tables 9.13, 9.15, 9.17, 9.20, and 9.22 illustrate the pairwise comparison results across the presentation conditions for user satisfaction measures. These tables are presented in relevant discussions below.

Ease of use. There was a significant difference in the ease of use ratings between the presentation conditions for all three supporting features: (see Table 9.12): policy ($H(3) = 33.14$, $p < .001$), creation suggestion ($H(3) = 49.74$, $p < .001$), and strength indicator ($H(3) = 34.68$, $p < .001$). In relation to all three supporting features, participants rated the ease of providing only one feature at a time (i.e. the *policy-only*,

suggestion-only, and *indicator-only* conditions) significantly lower than providing a combination of features regardless of the types of combinations. The pairwise comparison confirmed this difference (see Table 9.13).

Table 9.12 Mean (median) ratings of ease of use measures for policy, suggestion, and indicator across different presentation conditions

	<i>policy-only</i>	<i>suggestion-only</i>	<i>indicator-only</i>	<i>policy&suggestion</i>	<i>policy&indicator</i>	<i>suggestion&indicator</i>	<i>policy&suggestion&indicator</i>	<i>P value</i>
Policy	2.50 (2.00)	-	-	3.58 (4.00)	3.52 (4.00)	-	3.75 (4.00)	.000
Creation suggestion	-	2.10 (1.00)	-	3.58 (4.00)	-	3.66 (4.00)	3.75 (4.00)	.000
Strength indicator	-	-	2.46 (2.00)	-	3.52 (4.00)	3.66 (4.00)	3.75 (4.00)	.000

Table 9.13 Pairwise comparisons of ease of use measure ratings for policy, suggestion, and indicator across different presentation conditions

Ease of use		<i>policy-only</i>	<i>policy&suggestion</i>	<i>policy&indicator</i>	<i>policy&suggestion&indicator</i>
		Policy	<i>policy-only</i>	-	-52.01*
	<i>policy&suggestion</i>		-	3.95	-8.23
	<i>policy&indicator</i>			-	-12.19
	<i>policy&suggestion&indicator</i>				-
Creation suggestion		<i>suggestion-only</i>	<i>policy&suggestion</i>	<i>suggestion&indicator</i>	<i>policy&suggestion&indicator</i>
	<i>suggestion-only</i>	-	-63.39*	-67.09*	-70.73*
	<i>policy&suggestion</i>		-	-3.70	-7.34
	<i>suggestion&indicator</i>			-	-3.64
	<i>policy&suggestion&indicator</i>				-
Strength indicator		<i>indicator-only</i>	<i>policy&indicator</i>	<i>suggestion&indicator</i>	<i>policy&suggestion&indicator</i>
	<i>indicator-only</i>	-	-48.68*	-55.90*	-60.07*
	<i>policy&indicator</i>		-	-7.23	-11.39
	<i>suggestion&indicator</i>			-	-4.16
	<i>policy&suggestion&indicator</i>				-

Note. * denotes a significant pairwise comparison result, $p < .05$.

Annoyingness. There was a significant difference in the annoyingness ratings between the presentation conditions for all three supporting features: (see Table 9.14): policy ($H(3) = 51.51, p < .001$), creation suggestion ($H(3) = 45.17, p < .001$), and strength indicator ($H(3) = 24.43, p < .001$). In relation to all three supporting features,

participants perceived the provision of only one feature at a time (i.e. the policy-only, suggestion-only, and indicator-only conditions) as significantly high in annoyingness compared to the provision of a combination of features, regardless of the types of combinations. The pairwise comparison confirmed this difference (see Table 9.15).

Table 9.14 Mean (median) ratings of annoyingness measures for policy, suggestion, and indicator across different presentation conditions

	<i>policy-only</i>	<i>suggestion-only</i>	<i>indicator-only</i>	<i>policy&suggestion</i>	<i>policy&indicator</i>	<i>suggestion&indicator</i>	<i>policy&suggestion&indicator</i>	<i>p value</i>
Policy	2.41 (2.00)	-	-	3.91 (4.00)	3.71 (4.00)	-	4.02 (4.00)	.000
Creation suggestion	-	2.41 (2.00)	-	3.91 (4.00)	-	3.68 (4.00)	4.02 (4.00)	.000
Strength indicator	-	-	2.85 (3.00)	-	3.71 (4.00)	3.68 (4.00)	4.02 (4.00)	.000

Table 9.15 Pairwise comparisons of annoyingness measure ratings for policy, suggestion, and indicator across different presentation conditions

		<i>policy-only</i>	<i>policy&suggestion</i>	<i>policy&indicator</i>	<i>policy&suggestion&indicator</i>	
Annoyingness	Policy	<i>policy-only</i>	-	-67.86*	-58.28*	-75.56*
		<i>policy&suggestion</i>		-	9.58	-4.70
		<i>policy&indicator</i>			-	-14.28
		<i>policy&suggestion&indicator</i>				-
	Creation suggestion	<i>suggestion-only</i>	-	-64.23*	-55.33*	-68.87*
		<i>policy&suggestion</i>		-	-8.90	-4.64
		<i>suggestion&indicator</i>			-	-13.55
		<i>policy&suggestion&indicator</i>				-
	Strength indicator	<i>indicator-only</i>	-	-39.77*	-40.40*	-54.74*
		<i>policy&indicator</i>		-	-0.63	-14.97
		<i>suggestion&indicator</i>			-	-14.34
		<i>policy&suggestion&indicator</i>				-

Note. * denotes a significant pairwise comparison result, $p < .05$.

Helpfulness. There was a significant difference in the helpfulness ratings between the presentation conditions for all three supporting features: (see Table 9.16): policy (H

(3) = 29.03, $p < .001$), creation suggestion ($H(3) = 29.69$, $p < .001$), and strength indicator ($H(3) = 43.62$, $p < .001$). In relation to all three supporting features, ratings of the helpfulness of providing only one feature at a time (i.e. the *policy-only*, *suggestion-only*, and *indicator-only* conditions) were significantly lower than for the helpfulness of providing a combination of features, regardless of the types of combination. The pairwise comparison confirmed this difference (see Table 9.17).

Table 9.16 Mean (median) ratings of helpfulness measures for policy, suggestion, and indicator across different presentation conditions

	<i>policy-only</i>	<i>suggestion-only</i>	<i>indicator-only</i>	<i>policy & suggestion</i>	<i>policy & indicator</i>	<i>suggestion & indicator</i>	<i>policy & suggestion & indicator</i>	<i>p value</i>
Policy	3.16 (3.00)	-	-	3.65 (4.00)	4.00 (4.00)	-	3.81 (4.00)	.000
Creation suggestion	-	3.05 (3.00)	-	3.65 (4.00)	-	3.88 (4.00)	3.81 (4.00)	.000
Strength indicator	-	-	2.99 (3.00)	-	4.00 (4.00)	3.88 (4.00)	3.81 (4.00)	.000

Table 9.17 Pairwise comparisons of helpfulness measure ratings for policy, suggestion, and indicator across different presentation conditions

Helpfulness		<i>policy-only</i>	<i>policy & suggestion</i>	<i>policy & indicator</i>	<i>policy & suggestion & indicator</i>
		Policy	<i>policy-only</i>	-	-38.52*
		<i>policy & suggestion</i>	-	18.77	-6.61
		<i>policy & indicator</i>		-	-12.16
		<i>policy & suggestion & indicator</i>			-
Helpfulness		<i>suggestion-only</i>	<i>policy & suggestion</i>	<i>suggestion & indicator</i>	<i>policy & suggestion & indicator</i>
		Creation suggestion	<i>suggestion-only</i>	-	-44.81*
		<i>policy & suggestion</i>		-	-9.49
		<i>suggestion & indicator</i>		-	-2.79
		<i>policy & suggestion & indicator</i>			-
Helpfulness		<i>indicator-only</i>	<i>policy & indicator</i>	<i>suggestion & indicator</i>	<i>policy & suggestion & indicator</i>
		Strength indicator	<i>indicator-only</i>	-	-67.14*
		<i>policy & indicator</i>		-	9.44
		<i>suggestion & indicator</i>		-	-3.03
		<i>policy & suggestion & indicator</i>			-

Note. * denotes a significant pairwise comparison result, $p < .05$.

Clarity. There was no significant difference in the clarity ratings between the presentation conditions for all three features (see Table 9.18): policy ($H(3) = 2.74$, p

= .434), creation suggestion ($H(3) = 0.55, p = .908$), and strength indicator ($H(3) = 0.82, p = .845$).

Table 9.18 Mean (median) ratings of clarity measures for policy, suggestion, and indicator across different presentation conditions

	<i>policy-only</i>	<i>suggestion-only</i>	<i>indicator-only</i>	<i>policy&suggestion</i>	<i>policy&indicator</i>	<i>suggestion&indicator</i>	<i>policy&suggestion&indicator</i>	<i>p value</i>
Policy	3.84 (4.00)	-	-	3.96 (4.00)	4.09 (4.00)	-	3.96 (4.00)	n.s.
Creation suggestion	-	3.90 (4.00)	-	3.96 (4.00)	-	4.02 (4.00)	3.96 (4.00)	n.s.
Strength indicator	-	-	3.94 (4.00)	-	4.09 (4.00)	4.02 (4.00)	3.96 (4.00)	n.s.

Amount of detail. There was a significant difference in the amount of detail ratings between the presentation conditions for all three supporting features: (see Table 9.19): policy ($H(3) = 27.07, p < .001$), creation suggestion ($H(3) = 57.14, p < .001$), and strength indicator ($H(3) = 26.32, p < .001$). In relation to all three supporting features, participants rated the amount of detail presented in only one feature at a time (i.e. the *policy-only*, *suggestion-only*, and *indicator-only* conditions) significantly lower than the amount of detail presented within a combination of features, regardless of the types of combinations. The pairwise comparison confirmed this difference (see Table 9.20).

Table 9.19 Mean (median) ratings of amount of detail measures for policy, suggestion, and indicator across different presentation conditions

	<i>policy-only</i>	<i>suggestion-only</i>	<i>indicator-only</i>	<i>policy&suggestion</i>	<i>policy&indicator</i>	<i>suggestion&indicator</i>	<i>policy&suggestion&indicator</i>	<i>p value</i>
Policy	2.06 (2.00)	-	-	2.54 (3.00)	2.46 (3.00)	-	2.73 (3.00)	.000
Creation suggestion	-	1.68 (2.00)	-	2.54 (3.00)	-	2.44 (3.00)	2.73 (3.00)	.000
Strength indicator	-	-	2.01 (2.00)	-	2.46 (3.00)	2.44 (3.00)	2.73 (3.00)	.000

Table 9.20 Pairwise comparisons of amount of detail measure ratings for policy, suggestion, and indicator across different presentation conditions

		<i>policy- only</i>	<i>policy& suggestion</i>	<i>policy& indicator</i>	<i>policy&suggest ion&indicator</i>	
Amount of detail	Policy	<i>policy-only</i>	-	-39.77*	-31.26*	-53.95*
		<i>policy&suggestion</i>		-	8.51	-14.17
		<i>policy&indicator</i>			-	-22.68*
		<i>policy&suggestion&indicator</i>				-
	Creation suggestion		<i>suggesti on-only</i>	<i>policy& suggestion</i>	<i>suggestion &indicator</i>	<i>policy&suggest ion&indicator</i>
		<i>suggestion-only</i>	-	-63.07*	-56.90*	-76.46*
		<i>policy&suggestion</i>		-	6.18	-13.39
		<i>suggestion&indicator</i>			-	-19.57
	Strength indicator		<i>indicato r-only</i>	<i>policy& indicator</i>	<i>suggestion &indicator</i>	<i>policy&suggest ion&indicator</i>
		<i>indicator-only</i>	-	-33.03*	-34.96*	-55.97*
		<i>policy&indicator</i>		-	-1.93	-22.95
		<i>suggestion&indicator</i>			-	-21.01
	<i>policy&suggestion&indicator</i>				-	

Note. * denotes a significant pairwise comparison result, $p < .05$.

Confidence. There was a significant difference in the confidence ratings between the presentation conditions for only one feature (see Table 9.21): policy ($H(3) = 10.32$, $p = .016$). No significant difference was found for the other two features: creation suggestion ($H(3) = 2.82$, $p = .421$) and strength indicator ($H(3) = 4.35$, $p = .226$). In relation to the policy feature, participants felt less confident creating a password when only one feature was presented at a time (i.e. the *policy-only* condition) than in the other types of combination conditions (i.e. *policy&suggestion* and *policy&indicator*). The pairwise comparison confirmed this difference (see Table 9.22).

Table 9.21 Mean (median) ratings of confidence measures for policy, suggestion, and indicator across different presentation conditions

	<i>policy-only</i>	<i>suggestion-only</i>	<i>indicator-only</i>	<i>policy&suggestion</i>	<i>policy&indicator</i>	<i>suggestion&indicator</i>	<i>policy&suggestion&indicator</i>	<i>p value</i>
Policy	3.53 (4.00)	-	-	4.05 (4.00)	4.04 (4.00)	-	3.88 (4.00)	.016
Creation suggestion	-	3.83 (4.00)	-	4.05 (4.00)	-	3.75 (4.00)	3.88 (4.00)	n.s.
Strength indicator	-	-	3.70 (4.00)	-	4.04 (4.00)	3.75 (4.00)	3.88 (4.00)	n.s.

Table 9.22 Pairwise comparisons of confidence measure ratings for policy across different presentation conditions

Confidence	Policy	<i>policy-only</i>	<i>policy&suggestion</i>	<i>policy&indicator</i>	<i>policy&suggestion&indicator</i>
		<i>policy-only</i>	-	-31.27*	-30.72*
	<i>policy&suggestion</i>		-	0.55	13.40
	<i>policy&indicator</i>			-	12.85
	<i>policy&suggestion&indicator</i>				-

Note. * denotes a significant pairwise comparison result, $p < .05$.

9.3.3.1.2 Strength of Password

9.3.3.1.2.1 Password Characteristics

In this section, Tables 9.23, 9.25, 9.27, 9.29, and 9.31 summarise the between-participant results for all three supporting features. Tables 9.24, 9.26, 9.28, 9.30, and 9.32 illustrate the pairwise comparison results across the presentation conditions for password characteristics measures. These tables are presented in relevant discussions below.

Password length. There was a significant difference in password length between the presentation conditions for all three supporting features (see Table 9.23): policy ($H(3) = 12.38$, $p = .006$), creation suggestion ($H(3) = 21.98$, $p < .001$), and strength indicator ($H(3) = 15.77$, $p = .001$). In relation to the policy and strength indicator features, passwords created with the help of these features in combination with the creation

suggestion (i.e. the *policy&suggestion* and *suggestion&indicator* conditions) were significantly shorter in length. In terms of the creation suggestion feature, passwords created with all supporting features (i.e. the *policy&suggestion&indicator* condition) were significantly longer than those formulated with the other three presentation conditions. The pairwise comparison confirmed this difference (see Table 9.24).

Table 9.23 Mean (median) of password length measures for policy, suggestion, and indicator across different presentation conditions

	<i>policy-only</i>	<i>suggestion-only</i>	<i>indicator-only</i>	<i>policy&suggestion</i>	<i>policy&indicator</i>	<i>suggestion&indicator</i>	<i>policy&suggestion&indicator</i>	<i>p value</i>
Policy	12.79 (12.00)	-	-	11.79 (12.00)	12.82 (13.00)	-	13.06 (13.00)	.006
Creation suggestion	-	11.33 (11.00)	-	11.79 (12.00)	-	11.42 (11.00)	13.06 (13.00)	.000
Strength indicator	-	-	12.59 (12.00)	-	12.82 (13.00)	11.42 (11.00)	13.06 (13.00)	.001

Table 9.24 Pairwise comparisons of password length measures for policy, suggestion, and indicator across different presentation conditions

		<i>policy-only</i>	<i>policy&suggestion</i>	<i>policy&indicator</i>	<i>policy&suggestion&indicator</i>	
		Policy	<i>policy-only</i>	-	29.53*	-3.50
	<i>policy&suggestion</i>		-	-33.04*	-40.98*	
	<i>policy&indicator</i>			-	-7.94	
	<i>policy&suggestion&indicator</i>				-	
Password length	Creation suggestion	<i>suggestion-only</i>	<i>policy&suggestion</i>	<i>suggestion&indicator</i>	<i>policy&suggestion&indicator</i>	
		<i>suggestion-only</i>	-	-14.52	-2.83	-53.22*
		<i>policy&suggestion</i>		-	11.69	-38.70*
		<i>suggestion&indicator</i>			-	-50.39*
	<i>policy&suggestion&indicator</i>				-	
Strength indicator	Strength indicator	<i>indicator-only</i>	<i>policy&indicator</i>	<i>suggestion&indicator</i>	<i>policy&suggestion&indicator</i>	
		<i>indicator-only</i>	-	-12.63	-28.67*	-20.34
		<i>policy&indicator</i>		-	41.30*	-7.71
		<i>suggestion&indicator</i>			-	-49.01*
	<i>policy&suggestion&indicator</i>				-	

Note. * denotes a significant pairwise comparison result, $p < .05$.

Number of digits. There was a significant difference in the number of digits used in passwords between the presentation conditions in relation to only one feature (see Table 9.25): strength indicator ($H(3) = 9.26, p = .026$). No significant difference was found in connection with the other two features: policy ($H(3) = 6.75, p = .080$) and creation suggestion ($H(3) = 7.60, p = .055$). In relation to the strength indicator feature, passwords that were created when only one feature was provided at a time (i.e. the *indicator-only* condition) had more digits than those formulated in the other types of combination conditions (i.e. *policy&suggestion* and *policy&indicator*). The pairwise comparison confirmed this difference (see Table 9.26).

Table 9.25 Mean (median) number of digits measures for policy, suggestion, and indicator across different presentation conditions

	<i>policy-only</i>	<i>suggestion-only</i>	<i>indicator-only</i>	<i>policy&suggestion</i>	<i>policy&indicator</i>	<i>suggestion&indicator</i>	<i>policy&suggestion&indicator</i>	<i>p value</i>
Policy	2.84 (3.00)	-	-	3.37 (4.00)	2.63 (3.00)	-	3.08 (4.00)	n.s.
Creation suggestion	-	2.79 (2.00)	-	3.37 (4.00)	-	2.63 (3.00)	3.08 (4.00)	n.s.
Strength indicator	-	-	3.43 (4.00)	-	2.63 (3.00)	2.63 (3.00)	3.08 (4.00)	.026

Table 9.26 Pairwise comparisons of number of digits measures for indicator across different presentation conditions

Number of digits	Strength indicator	<i>indicator-only</i>	<i>policy&indicator</i>	<i>suggestion&indicator</i>	<i>policy&suggestion&indicator</i>
		<i>indicator-only</i>	-	29.38*	30.13*
	<i>policy&indicator</i>		-	0.75	-19.96
	<i>suggestion&indicator</i>			-	-20.70
	<i>policy&suggestion&indicator</i>				-

Note. * denotes a significant pairwise comparison result, $p < .05$.

Number of uppercase letters. There was a significant difference in the number of uppercase letters used in passwords between the presentation conditions for two features (see Table 9.27): creation suggestion ($H(3) = 12.65, p = .005$) and strength indicator ($H(3) = 11.12, p = .011$). No significant difference was found for policy (H

(3) = 3.57, $p = .312$). In relation to the creation suggestion feature, passwords created in the *policy&suggestion* and *policy&suggestion&indicator* conditions had significantly more uppercase letters than those formed in the other two conditions. With the strength indicator feature, passwords created when the strength indicator was combined with the suggestion statement (i.e. the *suggestion&indicator* condition) had fewer uppercase letters than those formulated in the other types of combination conditions. The pairwise comparison confirmed this difference (see Table 9.28).

Table 9.27 Mean (median) number of uppercase letters measures for policy, suggestion, and indicator across different presentation conditions

	<i>policy-only</i>	<i>suggestion-only</i>	<i>indicator-only</i>	<i>policy&suggestion</i>	<i>policy&indicator</i>	<i>suggestion&indicator</i>	<i>policy&suggestion&indicator</i>	<i>p value</i>
Policy	1.66 (1.00)	-	-	1.25 (1.00)	1.58 (1.00)	-	1.18 (1.00)	n.s.
Creation suggestion	-	0.73 (1.00)	-	1.25 (1.00)	-	0.74 (1.00)	1.18 (1.00)	.005
Strength indicator	-	-	1.30 (1.00)	-	1.58 (1.00)	0.74 (1.00)	1.18 (1.00)	.011

Table 9.28 Pairwise comparisons of number of uppercase letters measures for suggestion and indicator across different presentation conditions

Number of uppercase	Creation suggestion	<i>suggestion-only</i>	<i>policy&suggestion</i>	<i>suggestion&indicator</i>	<i>policy&suggestion&indicator</i>	
		-	-27.71*	-1.24	-32.50*	
Number of uppercase	Creation suggestion	<i>policy&suggestion</i>	-	26.48*	-4.78	
		<i>suggestion&indicator</i>	-	-	-31.26*	
		<i>policy&suggestion&indicator</i>	-	-	-	
	Strength indicator	<i>indicator-only</i>	-	-12.63	24.62*	-5.08
		<i>policy&indicator</i>	-	-	40.85*	11.14
<i>suggestion&indicator</i>		-	-	-	-29.70*	
<i>policy&suggestion&indicator</i>		-	-	-	-	

Note. * denotes a significant pairwise comparison result, $p < .05$.

Number of lowercase letters. There was a significant difference in the number of lowercase letters used in passwords between the presentation conditions for only one feature (see Table 9.29): policy ($H(3) = 8.05$, $p = .045$). No significant difference was

found for the other two features: creation suggestion ($H(3) = 4.37, p = .224$) and strength indicator ($H(3) = 1.96, p = .581$). Within the policy feature, passwords created in the *policy&suggestion* condition had significantly fewer lowercase letters than those formed in the other conditions (i.e. *policy&indicator* and *policy&suggestion&indicator*). The pairwise comparison confirmed this difference (see Table 9.30).

Table 9.29 Mean (median) number of lowercase letters measures for policy, suggestion, and indicator across different presentation conditions

	<i>policy-only</i>	<i>suggestion-only</i>	<i>indicator-only</i>	<i>policy&suggestion</i>	<i>policy&indicator</i>	<i>suggestion&indicator</i>	<i>policy&suggestion&indicator</i>	<i>p value</i>
Policy	6.64 (7.00)	-	-	5.67 (6.00)	7.16 (7.00)	-	6.88 (6.00)	.045
Creation suggestion	-	6.20 (6.00)	-	5.67 (6.00)	-	6.38 (6.00)	6.88 (6.00)	n.s.
Strength indicator	-	-	6.60 (7.00)	-	7.16 (7.00)	6.38 (6.00)	6.88 (6.00)	n.s.

Table 9.30 Pairwise comparisons of number of lowercase letters measures for policy across different presentation conditions

Number of	Policy	<i>policy-only</i>	<i>policy&suggestion</i>	<i>policy&indicator</i>	<i>policy&suggestion&indicator</i>
		<i>policy-only</i>	-	22.02	-11.43
<i>policy&suggestion</i>			-	-33.44*	-25.96*
<i>policy&indicator</i>				-	-7.48
<i>policy&suggestion&indicator</i>					-

Note. * denotes a significant pairwise comparison result, $p < .05$.

Number of symbols. There was a significant difference in the number of symbols used in passwords among the presentation conditions for only one feature (see Table 9.31): strength indicator ($H(3) = 11.61, p = .009$). No significant difference was found for the other two features: policy ($H(3) = 4.19, p = .242$) and creation suggestion ($H(3) = 6.38, p = .095$). Within the strength indicator feature, passwords created when the strength indicator was combined with the suggestion statement (i.e.

suggestion&indicator) had significantly more symbols than those formulated in the other conditions. The pairwise comparison confirmed this difference (see Table 9.32)

Table 9.31 Mean (median) number of symbols measures for policy, suggestion, and indicator across different presentation conditions

	<i>policy-only</i>	<i>suggestion-only</i>	<i>indicator-only</i>	<i>policy&suggestion</i>	<i>policy&indicator</i>	<i>suggestion&indicator</i>	<i>policy&suggestion&indicator</i>	<i>p value</i>
Policy	0.82 (1.00)	-	-	1.09 (1.00)	1.09 (1.00)	-	0.80 (1.00)	n.s.
Creation suggestion	-	1.07 (1.00)	-	1.09 (1.00)	-	1.02 (1.00)	0.80 (1.00)	n.s.
Strength indicator	-	-	0.67 (1.00)	-	1.09 (1.00)	1.02 (1.00)	0.80 (1.00)	.009

Table 9.32 Pairwise comparisons of number of symbols measures for indicator across different presentation conditions

Number of symbols Strength indicator		<i>indicator-only</i>	<i>policy&indicator</i>	<i>suggestion&indicator</i>	<i>policy&suggestion&indicator</i>
	<i>indicator-only</i>		-	-21.21	-36.82*
<i>policy&indicator</i>			-	-15.61	7.36
<i>suggestion&indicator</i>				-	-22.97
<i>policy&suggestion&indicator</i>					-

Note. * denotes a significant pairwise comparison result, $p < .05$.

Number of password character classes. Counting the number of different character classes that occurred in passwords, there were significant differences in the number of character classes used in passwords in all seven presentation conditions: *policy-only* ($\chi^2(3) = 46.47$, $p < .001$), *suggestion-only* ($\chi^2(3) = 23.54$, $p < .001$), *indicator-only* ($\chi^2(3) = 19.10$, $p < .001$), *policy&suggestion* ($\chi^2(2) = 14.00$, $p = .001$), *policy&indicator* ($\chi^2(3) = 22.71$, $p < .001$), *suggestion&indicator* ($\chi^2(2) = 24.86$, $p < .001$), and *policy&suggestion&indicator* ($\chi^2(2) = 11.38$, $p = .003$). The distribution of the password character classes across the presentation conditions is presented in Figure 9.15. The *policy&suggestion* condition had the highest percentage of passwords using all four character classes (54.39%); it is followed by the *policy-only* condition (52.94%).

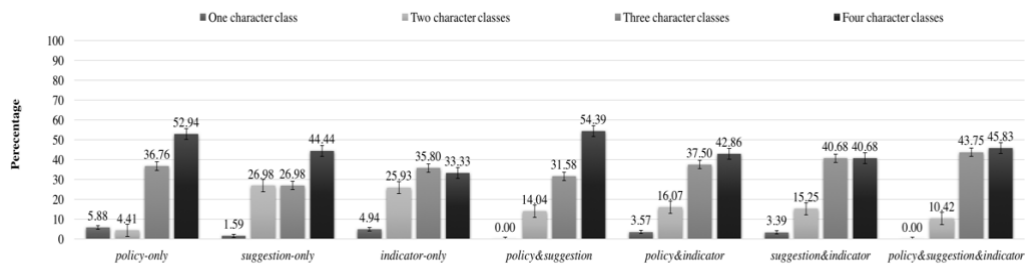


Figure 9.15 Percentage of password character classes for policy, suggestion, and indicator across different presentation conditions

Policy compliance. There was a significant difference in the number of passwords that followed the provided policy in only two presentation conditions: *policy-only* ($\chi^2(1) = 11.53$, $p = .001$) and *policy&suggestion&indicator* ($\chi^2(1) = 8.33$, $p = .004$). No significant difference was found in the other two conditions: *policy&suggestion* ($\chi^2(1) = 0.86$, $p = .354$) and *policy&indicator* ($\chi^2(1) = 2.57$, $p = .109$). The distribution of policy compliance across the three types of combination conditions is presented in Figure 9.16. The percentage of passwords that followed the policy was above 70% for both the *policy-only* and *policy&suggestion&indicator* conditions.

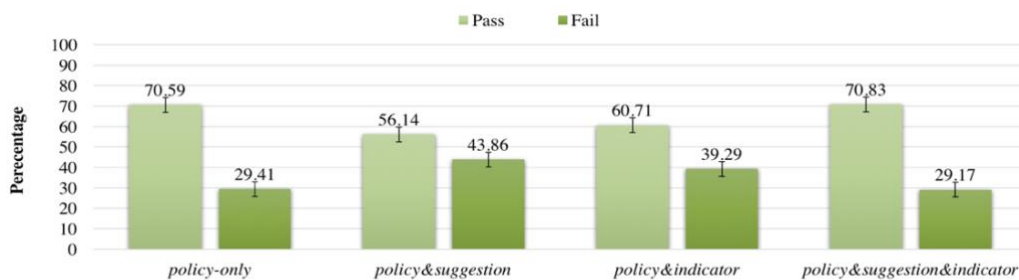


Figure 9.16 Percentage of policy compliance across different presentation conditions

Suggestion compliance. There was a significant difference in the number of passwords that followed the provided suggestion in two presentation conditions: *policy&suggestion* ($\chi^2(1) = 7.74$, $p = .005$) and *suggestion&indicator* ($\chi^2(1) = 8.97$, $p = .003$). No significant difference was found in the *suggestion-only* ($\chi^2(1) = 1.92$, $p = .166$) and *policy&suggestion&indicator* ($\chi^2(1) = 2.08$, $p = .149$) conditions. The distribution of suggestion compliance across the three types of combination conditions

is presented in Figure 9.17. The *suggestion&indicator* condition had the highest percentage of passwords that followed the suggestion (69.49%); the *policy&suggestion* condition had the next highest percentage (68.42%).

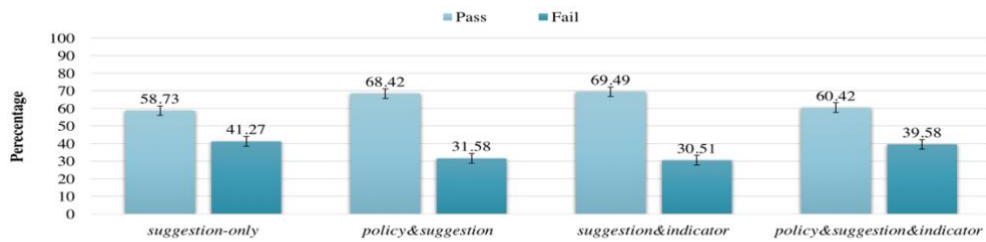


Figure 9.17 Percentage of suggestion compliance across different presentation conditions

Symbols provision. There was a significant difference in the number of passwords that included the provided symbols in only one presentation condition: *suggestion-only* ($\chi^2(1) = 9.92, p = .002$). No significant difference was found in the other three conditions *policy&suggestion* ($\chi^2(1) = 1.42, p = .233$), *suggestion&indicator* ($\chi^2(1) = 2.86, p = .091$) and *policy&suggestion&indicator* ($\chi^2(1) = 1.33, p = .248$). The percentages of passwords that included at least one of the symbols are presented across the three types of combination conditions in Figure 9.18. Overall, 69.84% of passwords included at least one of the symbols in the *suggestion-only* condition.

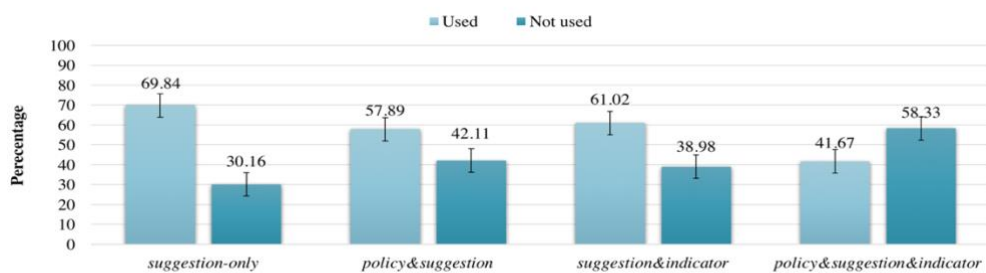


Figure 9.18 Percentage of passwords that included the given symbols across different presentation conditions

Password strength score. There was a significant difference in the password strength scores between the different presentation conditions for the password strength feature ($H(3) = 12.96, p = .005$). Passwords created with the *suggestion&indicator* ($M =$

72.28, $Mdn = 72.27$) condition were weaker than those formed in the other two conditions: *policy&indicator* ($M = 78.47$, $Mdn = 78.84$) and *policy&suggestion&indicator* ($M = 80.71$, $Mdn = 80.78$). There was also a significant difference in the password strength score between the *indicator-only* ($M = 75.46$, $Mdn = 73.30$) and *policy&suggestion&indicator* conditions. The pairwise comparison confirmed this difference, as Table 9.33 indicates. The distribution of password strength levels across the different conditions is presented in Figure 9.19. The highest percentage of *strong* passwords was created in the *policy&suggestion&indicator* condition (83.33%), which was followed by the *policy&indicator* condition (80.36%).

Table 9.33 Pairwise comparisons of strength score measures for indicator across different presentation conditions

Password strength score	Strength indicator	<i>indicat</i>	<i>policy&</i>	<i>suggestion&i</i>	<i>policy&suggest</i>
		<i>or-only</i>	<i>indicator</i>	<i>ndicator</i>	<i>ion&indicator</i>
	<i>indicator-only</i>	-	-18.59	15.57	-28.92*
	<i>policy&indicator</i>		-	34.16*	-10.33
	<i>suggestion&indicator</i>			-	-44.49*
	<i>policy&suggestion&indicator</i>				-

Note. * denotes a significant pairwise comparison result, $p < .05$.

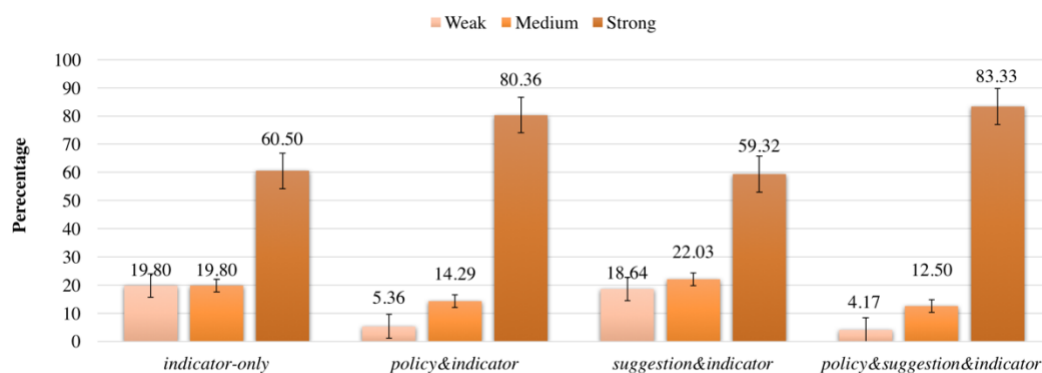


Figure 9.19 Percentage of password strength levels across different presentation conditions

9.3.3.1.2.2 Password Guessability

The distribution of password guessability across different presentation conditions is presented in Figure 9.20. The results revealed that there was a significant difference

in only four presentation conditions: *policy-only* ($x^2(1) = 9.38$, $p = .002$), *indicator-only* ($x^2(1) = 7.72$, $p = .005$), *policy&suggestion* ($x^2(1) = 6.13$, $p = .013$), and *policy&indicator* ($x^2(1) = 9.53$, $p = .002$). No significant difference was found in the *suggestion-only* ($x^2(1) = 0.40$, $p = .529$), *suggestion&indicator* ($x^2(1) = 0.02$, $p = .896$), and *policy&suggestion&indicator* ($x^2(1) = 1.88$, $p = .170$) conditions. The *policy&indicator* condition had the highest percentage of passwords that were not guessable (76.47%).

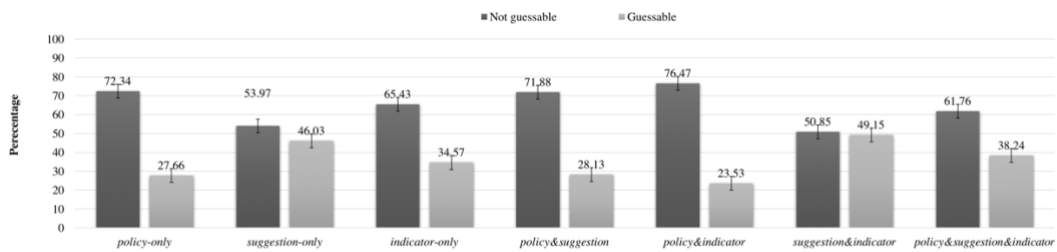


Figure 9.20 Percentage of password guessability across different presentation conditions

9.3.3.2 Password Recall

The distribution of the participants who returned to take part in the Part II recall task was as follows: 47 in *policy-only*, 43 in *suggestion-only*, 51 in *indicator-only*, 37 in *policy&suggestion*, 44 in *policy&indicator*, 33 in *suggestion&indicator*, and 33 in *policy&suggestion&indicator*. Based on this sample, this section presents the results in relation to PCS usability when users recall their passwords.

9.3.3.2.1 Efficiency Measures

Recall time. Table 9.34 summarises the between-participant results for all three supporting features; Table 9.35 illustrates the pairwise comparison results across the presentation conditions for the recall time.

There was a significant difference in recall time between the presentation conditions for all three features (see Table 9.34): policy ($H(3) = 11.17$, $p = .011$), creation suggestion ($H(3) = 12.19$, $p = .007$), and strength indicator ($H(3) = 15.06$, $p = .002$).

In relation to all three supporting features, participants spent less time recalling passwords created with only one feature at a time (i.e. the *policy-only*, *suggestion-only*, and *indicator-only* conditions) than they did recalling those created with a combination of features. The pairwise comparison confirmed this difference (see Table 9.35).

Table 9.34 Mean (median) of recall time measures for policy, suggestion, and indicator across different presentation conditions

	<i>policy-only</i>	<i>suggestion-only</i>	<i>indicator-only</i>	<i>policy&suggestion</i>	<i>policy&indicator</i>	<i>suggestion&indicator</i>	<i>policy&suggestion&indicator</i>	<i>p value</i>
Policy	22.45 (21.00)	-	-	29.65 (25.00)	27.36 (24.50)	-	34.05 (28.00)	.011
Creation suggestion	-	21.57 (21.00)	-	29.65 (25.00)	-	24.73 (22.00)	34.05 (28.00)	.007
Strength indicator	-	-	20.87 (19.00)	-	27.36 (24.50)	24.73 (22.00)	34.05 (28.00)	.002

Table 9.35 Pairwise comparisons of recall time measures for policy, suggestion, and indicator across different presentation conditions

		<i>policy-only</i>	<i>policy&suggestion</i>	<i>policy&indicator</i>	<i>policy&suggestion&indicator</i>
		Recall time	<i>policy-only</i>	-	-26.79*
	<i>policy&suggestion</i>		-	8.99	-13.92
	<i>policy&indicator</i>			-	-13.92
	<i>policy&suggestion&indicator</i>				-
		<i>suggestion-only</i>	<i>policy&suggestion</i>	<i>suggestion&indicator</i>	<i>policy&suggestion&indicator</i>
	<i>suggestion-only</i>	-	-25.04*	-10.06	-30.18*
	<i>policy&suggestion</i>		-	14.97	-5.15
	<i>suggestion&indicator</i>			-	-20.12
	<i>policy&suggestion&indicator</i>				-
		<i>indicator-only</i>	<i>policy&indicator</i>	<i>suggestion&indicator</i>	<i>policy&suggestion&indicator</i>
	<i>indicator-only</i>	-	-24.58*	-15.31	-38.64*
	<i>policy&indicator</i>		-	-9.28	-14.06
	<i>suggestion&indicator</i>			-	-23.33*
	<i>policy&suggestion&indicator</i>				-

Note. * denotes a significant pairwise comparison result, $p < .05$.

9.3.3.2.2 Effectiveness Measures

Accuracy. Figure 9.21 presents the percentage of password recall accuracy for all of the presentation conditions. There was a significant difference in password recall accuracy in only two presentation conditions: *policy-only* ($x^2(1) = 4.79$, $p = .029$) and *indicator-only* ($x^2(1) = 4.41$, $p = .036$). No significant difference was found in the *suggestion-only* ($x^2(1) = 2.81$, $p = .093$), *policy&suggestion* ($x^2(1) = 0.24$, $p = .622$), *policy&indicator* ($x^2(1) = 0.09$, $p = .763$), *suggestion&indicator* ($x^2(1) = 0.03$, $p = .862$), and *policy&suggestion&indicator* ($x^2(1) = 0.76$, $p = .384$) conditions. The highest percentage of successful recall was in the *policy-only* condition (34.04%), which was followed by the *indicator-only* condition (32.43%).

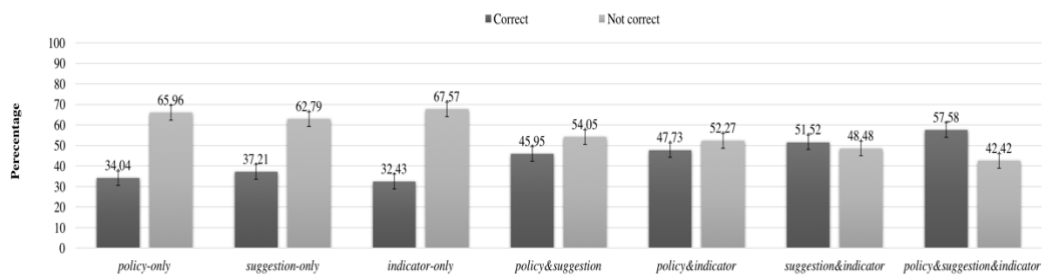


Figure 9.21 Percentage of password recall accuracy across different presentation conditions

9.3.3.2.3 User Satisfaction Measures

Confidence. Table 9.36 summarises the between-participant results for all three supporting features; Table 9.37 illustrates the pairwise comparison results among the presentation conditions for the confidence measure.

There was a significant difference in the ratings of confidence in recalling the correct password between the presentation conditions in all three features (see Table 9.36): policy ($H(3) = 16.64$, $p = .001$), creation suggestion ($H(3) = 12.95$, $p = .005$), and strength indicator ($H(3) = 16.30$, $p = .001$). In relation to all three supporting features, participants felt less confident recalling their correct password when only one feature was presented (i.e. the *policy-only*, *suggestion-only*, and *indicator-only* conditions)

than they did when provided with a combination of features. The pairwise comparison confirmed this difference (see Table 9.37).

Table 9.36 Mean (median) ratings of confidence measures for policy, suggestion, and indicator across different presentation conditions

	<i>policy-only</i>	<i>suggestion-only</i>	<i>indicator-only</i>	<i>policy&suggestion</i>	<i>policy&indicator</i>	<i>suggestion&indicator</i>	<i>policy&suggestion&indicator</i>	<i>p value</i>
Policy	2.64 (2.00)	-	-	3.51 (4.00)	3.68 (4.00)	-	3.55 (4.00)	.001
Creation suggestion	-	2.72 (3.00)	-	3.51 (4.00)	-	3.79 (4.00)	3.55 (4.00)	.005
Strength indicator	-	-	2.76 (3.00)	-	3.68 (4.00)	3.79 (4.00)	3.55 (4.00)	.001

Table 9.37 Pairwise comparisons of confidence measure ratings for policy, suggestion, and indicator across different presentation conditions

		<i>policy-only</i>	<i>policy&suggestion</i>	<i>policy&indicator</i>	<i>policy&suggestion&indicator</i>
Confidence	Policy	<i>policy-only</i>	-	-29.29*	-35.27*
		<i>policy&suggestion</i>		-	-5.99
		<i>policy&indicator</i>			-
		<i>policy&suggestion&indicator</i>			
	Creation suggestion	<i>suggestion-only</i>	-	-22.93*	-31.90*
		<i>policy&suggestion</i>		-	-8.96
		<i>suggestion&indicator</i>			-
		<i>policy&suggestion&indicator</i>			
	Strength indicator	<i>indicator-only</i>	-	-30.87*	-34.73*
		<i>policy&indicator</i>		-	-3.86
		<i>suggestion&indicator</i>			-
		<i>policy&suggestion&indicator</i>			

Note. * denotes a significant pairwise comparison result, $p < .05$.

9.3.4 Users' Common Password Creation and Recall Practices

Participants reported having an average of approximately 12.43 password-protected accounts (standard deviation = 16.71) and approximately 8.60 passwords (standard deviation = 13.45). However, most said they used either the same password (148,

67.27%) or slightly different passwords (143, 65%) for multiple accounts. When participants were asked about their actual behaviour in the current study, on average very few reported using a reused password (25, 11.36%) or modified version (32, 14.55%); most said they had created an entirely new password (163, 74.01%). This result might be due to the fact that participants were instructed to try their best to create a new password for this study.

Regarding password change frequency, most participants (88, 40.00%) reported changing their passwords every three to six months. Very few (8, 3.64%) had never changed them.

Many participants (88, 40.00%) described themselves as being very knowledgeable about what makes a secure password. Moreover, participants noted that secure passwords should be long, have a combination of different character classes, and not be based on personal information that might make them easy to guess. In this study, participants had the same perceptions of what makes a password secure as found in previous Studies 4, 5, and 6. Most participants (95, 43.18%) felt very confident about the strength of their most complicated password, and very few (2, 0.91%) did not feel confident at all.

Regarding password creation instructions, half of the participants (110, 50.00%) indicated that they 'always' read them. However, the participants explained the circumstances when they did not do so as (1) managed to create a password on their first attempt; (2) were familiar with instructions on the website, seeing as they were changing an existing password and not creating a new one; (3) were in a hurry; (4) the instructions were too lengthy or invisible in the PCS; or (5) the instructions were associated with low-value accounts. Again, the same circumstances were reported in previous Studies 4, 5, and 6.

Few participants (32, 14.55%) reported having a previous negative experience during the password creation process. They explained their frustration as follows: (1) the PCS enforced a very strict password policy that was associated with a low-value account,

(2) it took a long time to comply with the password policy, (3) a chosen password was not allowed even though it complied with the policy seeing as it was similar to a password the participant had used previously for the same account, and (4) the PCS implemented a CAPTCHA¹¹ that was very difficult to read. Interestingly, the first three situations were also mentioned in previous studies 4 and 6 whereas the last one was reported only in the current study.

In relation to password management strategies, the current study's participants cited the same strategies as reported in previous Studies 4, 5, and 6. They mentioned writing passwords down, using a password manager, reusing the same passwords for multiple accounts, employing different variations of one password, relying on their memory, and choosing easy-to-remember passwords. Participants reported different ways of keeping passwords safe when they choose to write them down, such as using notepads, sticky notes, and encrypted files on their computer. In this study, 34 participants (23.13%) reported writing their passwords down when asked about their password management behaviour.

9.4 Discussion

This study investigated the combined effects of providing more than one supporting feature at a time to users. To this end, the best design for each supporting feature as found in Study 6 (see Chapter 8) was used to design four types of combinations of supporting features (i.e. *policy&suggestion*, *policy&indicator*, *suggestion&indicator*, and *policy&suggestion&indicator*). This study consisted of two parts: password creation and password recall. Data from a total of 220 users was analysed in Part I,

¹¹ CAPTCHA: Completely Automated Public Turing test to tell Computers and Humans Apart.

while 147 returned for Part II. The first and second research questions (RQ1 and RQ2) examined the differences among various types of combinations when participants created and recalled passwords. The third and fourth research questions (RQ3 and RQ4) compared the provision of combined supporting features to the baseline for both password creation and recall. The password creation was examined in terms of PCS usability and password strength, whereas the password recall was examined in terms of only the usability of PCSs.

Regarding password creation, the efficiency of the PCSs and users' satisfaction with them were both used to determine the PCS usability levels. Password characteristics and guessability were used to examine the strength of the created passwords.

The current study found that the different types of combinations of supporting features had an effect on the PCS efficiency and password characteristics but not on user satisfaction and password guessability when participants created passwords. On the other hand, the results did not show an effect of the different types of combinations of supporting features on PCS efficiency or effectiveness and user satisfaction when participants recalled their passwords.

In terms of password creation, the four types of combinations affected PCS usability and password strength in different ways. The results revealed that when the policy statement or strength indicator was combined with the creation suggestion statement, the PCSs emerged as the most efficient design in relation to password creation time and keystrokes (see Table 9.3). Furthermore, each type of combination affected password characteristics differently (see Table 9.6): (1) participants created longer passwords when the PCS provided them with all three supporting features or the policy statement combined with the strength indicator; (2) participants chose passwords with a high number of digits but low number of lowercase letters when the PCS presented them the policy statement in combination with the creation suggestion, and finally (3) participants created passwords with a low number of uppercase letters when the PCS included both password strength and the creation suggestion statement. It thus seems

that combining the creation suggestion statement with another supporting feature affects PCS efficiency positively at the expense of password characteristics. This negative effect on password characters could be reduced by including all three supporting features in PCSs. Although presenting both policy and creation suggestion statements had a negative effect on password characteristics, it did improve both the number of password character classes and password guessability.

The results of the current study also revealed no effect on the level of user satisfaction between the types of combination, similar to findings from Study 6 (see Chapter 8) when one supporting feature was presented. This was somewhat surprising, seeing as one would expect that providing an extra supporting feature during the password creation process would improve user satisfaction, but this was not true. As mentioned in Study 6, showing no effect on the user satisfaction level might be due to the chosen policy and creation suggestion statements.

Another important finding from the present study was that providing a combination of supporting features (regardless of the types of combinations) affected PCS usability and password strength when participants created passwords but did not influence PCS usability when they recalled these passwords. In terms of password creation, providing a combination of supporting features negatively affected PCS efficiency but positively affected both user satisfaction and password strength. These results were consistent with those found in Study 6 (see Chapter 8). In relation to password recall, providing a combination of supporting features negatively affected the recall time.

The best presentation of each supporting feature found in Study 6 was used to design different combinations of features in the present study. A comparison between the individual and combined presentations of supporting features (Study 6 and the present study, respectively) yielded interesting findings. However, the results of these comparisons must be interpreted with caution given that the participants in Study 6 were exposed to five password creation tasks, which was more than twice the number of tasks used in the present study. These results are discussed next.

In general, the results of this comparison revealed that different presentations for all supporting features had an effect on user satisfaction but not on PCS efficiency, password characteristics, or password guessability when participants created passwords. However, these results also demonstrated that different presentations for all supporting features had an effect on PCS efficiency and user satisfaction but not on recall accuracy when participants recalled their passwords.

In terms of password creation, for all supporting features participants had lower user satisfaction when they were provided with only one feature (apart from clarity and confidence) than when they were presented a combination of features, regardless of the types of combinations. Interestingly, providing participants a combination of supporting features did not have an effect on PCS efficiency, except in relation to the number of keystrokes for the password strength indicator feature. Regarding password characteristics, most of the impact on password characteristics was associated with the password strength indicator feature. Depending on the type of supporting feature with which the strength indicator was combined, the effect varied. For example, when the strength indicator was combined with the policy statement, passwords included a high number of uppercase letters; when it was instead combined with the suggestion statement, passwords featured a high number of symbols. Furthermore, providing the policy statement combined with the strength indicator increased the level of non-guessable passwords in comparison to presenting these features individually.

In relation to password recall, providing only one supporting feature (regardless of feature) affected participants' recall times and confidence more negatively than presenting a combination of features. This was true regardless of the types of combinations.

The results of the present study must be interpreted with caution considering the limitations discussed in Studies 4 and 6 (see Chapters 6 and 8, respectively). One limitation is that an online study conducted to understand password behaviour lacks ecological validity, although the author implemented different ways to increase

ecological validity (e.g. using a scenario). The results showed nearly three quarters of the participants (74.01%) reported creating an entirely new password and nearly two thirds of the passwords created (62.89%) were not guessed by the cracking approaches. These percentages suggest that participants took the scenario and the password creation task seriously. Furthermore, 23.13% of participants who returned for Part II of the study reported writing their passwords down, which suggests that they behaved in the same way that they normally would when managing their passwords. It is crucial to note that self-reporting tends to be unreliable sometime. Finally, the general consistency in the outcomes of Studies 4, 5, and 6 and the present study suggests that the findings are valid.

9.5 Conclusions

To conclude, this study investigated the effects that presenting more than one supporting feature to users has on both these individuals and their passwords. Its main finding suggests that different combinations of supporting feature had some effects on PCSs usability and password strength. More precisely, effects were found on both the efficiency of PCSs and password characteristics but not on password guessability or user satisfaction. A further comparison was undertaken to determine whether presenting individual features affected the level of user satisfaction in comparison to presenting combined features. The findings revealed that having more than one supporting feature in the PCS improved user satisfaction. Another important result is that the presence of supporting features does improve both PCS usability and password strength, which was in line of the outcomes of Study 6.

***Phase 3: Proposing Usability Heuristics and
Guidelines for PCSs***

Chapter 10

Password Creation System Heuristics and Guidelines: Development and Evaluation – *Study 8*

10.1 Introduction

It is very useful to have appropriate usability heuristics and guidelines when conducting usability evaluations of interactive systems or designing them. The outcomes of Study 1 (see Chapter 3) showed PCSs are small interactive systems that can provide three main supporting features: password policy, password creation suggestions, and password strength indicators. Each of these features has its particular characteristics. The results from the usability evaluations (Studies 2 and 3, see Chapters 4 and 5, respectively) revealed that the distribution of the usability problems varied between the PCSs which were investigated. In these studies (2 and 3), the overlap between the usability problems found by the experts and users was small. Unexpectedly, the experts' performance was a disappointment as they missed 31.4% of the usability problems encountered by users in such small systems. Thus, it is important to support evaluators with an appropriate set of heuristics that could guide them through the evaluation of PCSs. To the best of the author's knowledge, the currently available sets of usability heuristics (e.g. Nielsen's heuristics and those developed by Petrie and Power) do not cover the diverse range of problems associated with password creation that were encountered by the experts and users. Hence, a set of heuristics and guidelines specifically for PCSs should be constructed to help evaluators and developers in their tasks.

With the evolution of interactive systems, researchers have developed new sets of usability heuristics for specific domains. Generally, two approaches have been adopted to develop new usability heuristics: bottom-up and top-down approaches (Jaferian, Hawkey, Sotirakopoulos, Velez-Rojas, & Beznosov, 2014). The majority of existing usability heuristics use the latter approach. The bottom-up approach relies on qualitative and real-world data by analysing the characteristics of appropriate systems, whereas the top-down approach relies on expert knowledge, by using relevant theories and existing heuristics. Examples of existing heuristics using the bottom-up approach are those developed for highly interactive websites (Petrie & Power, 2012) and video games (Pinelle, Wong, & Stach, 2008). Heuristics developed using the top-down approach include those for ambient displays (Mankoff et al., 2003) and shared visual workspaces (Baker, Greenberg, & Gutwin, 2002). The bottom-up approach was employed in the present study because it is grounded in real-world data reflecting real users' usability problems. Thus, the chance of missing information about usability problems related to the specific domain is hopefully minimised, as is the chance of including information about problems that users does not actually experience.

The present study aimed to develop a set of heuristics for evaluators and guidelines for developers to guide the evaluation and development of PCSs. Furthermore, it aimed to conduct an initial evaluation of the proposed heuristics. However, due to time and resource constraints, the present author only evaluated the heuristics and further work should evaluate the guidelines. To address these aims, the study consisted of two parts. Part I (see Section 10.2) covered the development process of the proposed set of heuristics (henceforth *PassHeuristics*) and guidelines (henceforth *PassGuidelines*), and Part II (see Section 10.3) evaluated the effectiveness of the heuristics. Part I comprised a content analysis of the categorisation of usability problems encountered by users (Study 3, see Chapter 5). In Part II, nine usability professionals were asked to use the heuristics in the evaluation of four mock-up PCSs to evaluate the *PassHeuristics* themselves. The following research question was formulated to address this aim:

RQ 3. Do evaluators find the new proposed heuristics, PassHeuristics, easy to use, clear, and useful in the evaluation of PCSs?

10.2 Development of the *PassHeuristics* and *PassGuidelines*

The development of the proposed heuristics and guidelines was the first part of this study. The starting point was the categorisation of the usability problems encountered by users (outcomes from the first phase of this research), which helped in identifying the information to include in the heuristics. At the same time, the results from the user experimental studies (outcomes from the second phase of this research) were used to feed in details needed for each heuristic that relate to the design guidelines, where applicable. For example, the usability evaluation study highlighted the lack of password strength indicator in PCSs as a problem (heuristic), whereas the experimental studies identified how the indicator should be presented and designed in terms of colour-scheme and media presentation (guideline). Each item in the proposed set of heuristics and guidelines was supported by both user perception and performance data. After that, a thorough review was conducted and feedback was obtained from a usability professional on several versions of the proposed heuristics and guidelines. Changes were made until a final satisfactory version emerged.

Section 10.2.1 discusses the data sources used in the development process; Section 10.2.2 presents the first version of the heuristics and guidelines, and the reasons for inclusion of particular items; then Section 10.2.3 illustrates the review process; and Section 10.2.4 presents the final version of the heuristics and guidelines.

10.2.1 Data Sources

The development of the heuristics started from the content analysis of usability problems encountered by users in Study 3 (see Section 5.3.2, Chapter 5). This results in 654 instances of usability problems and 81 distinct usability problems were included in the content analysis, falling into six main categories and 32 subcategories.

To extract information for the heuristics, several considerations were taken into account: (1) the number of users who identified the usability problem, (2) the number

of PCSs in which the problems were encountered, (3) the number of problems (both instances and distinct), and finally (4) the mean severity ratings of the problem. As a result, the first version of the *PassHeuristics* contained eight items. Figure 10.1 shows the percentages of the usability problems (instances and distinct) that were covered or not by this version of the *PassHeuristics*.

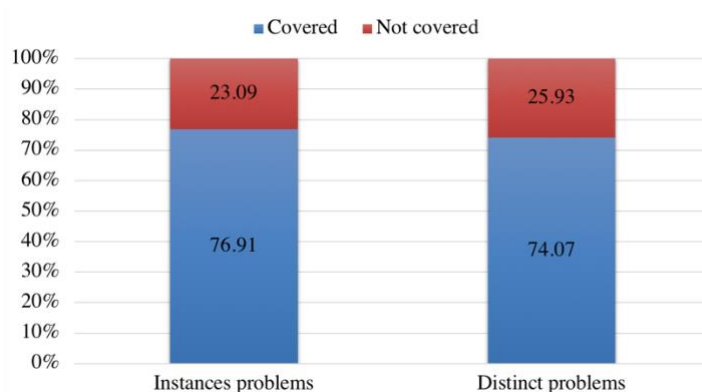


Figure 10.1 Proportion (%) of covered and not covered usability problems in the *PassHeuristics*

The proposed heuristics addressed 76.91% (503/654) of the instances and 74.07% (60/81) of the distinct usability problems. For each item included in the *PassHeuristics*, supporting evidence from user perception (best and worst practices of PCSs) and performance data (efficiency, perceived usability, and password strength) were used to create the *PassGuidelines*. Table 10.1 shows the source and supporting data for each heuristic.

Table 10.1 Source and supporting data for each item in the *PassHeuristics* and *PassGuidelines*

Heuristic/ Guideline	Phase 1		Phase 2		
	Study 3 Usability problems	Study 3 User perception	Study 5	Study 6 User performance	Study 7
#1	Source	Supporting		Supporting	
#2	Source	Supporting	Supporting		
#3	Source	Supporting			
#4	Source	Supporting			
#5	Source	Supporting		Supporting	
#6	Source	Supporting			
#7	Source	Supporting			
#8				Source	Source

However, 23.09% (151/654) of the instances and 25.93% (21/81) of the distinct usability problems were not covered in the *PassHeuristics* as they did not meet the

above four criteria. These problems not covered were related to: (1) querying the security of the policy, creation suggestion, and the PCS itself; (2) feedback and error message provided in the PCSs; and finally (3) overall presentation of the PCS. The following are examples of the problems not covered:

- *I think 4 is a bit small for a password and 60 characters is also an awful a lot in my opinion. It doesn't seem too safe! (P24);*
- *It is reacting a bit slowly (P16); and*
- *I expected this blue bar to go red when I did something wrong - it's not clear and this is not the commonly used method to show something is selected (P08).*

10.2.2 First Version of the *PassHeuristics* and *PassGuidelines*

The eight items developed in the first version of the proposed heuristics and guidelines and the rationale for their inclusion are presented below. The heuristics are shown in bold, whereas the design guidelines are shown underneath each heuristic in bullet-point.

Heuristic and Guideline #1

Provide the password policy before the users start entering their password and keep it available to them during and after password entry. Provide password creation suggestions only once password entry has begun and keep them available during and after password entry.

- The system should always provide the user with a statement of the password policy as soon as he lands on the page.
- The system may also provide password creation suggestions as soon as the user starts typing his password.
- In both cases, the information should remain visible throughout the password creation process.

The first heuristic and guideline concerned the provision and timing of presentation for both the statements of policy and creation suggestion. It was developed mainly from the usability problems users encountered in Study 3 and supported by evidence from user perception (Study 3) and user performance (Study 6). Table 10.2 illustrates the components of the first heuristic and guideline and the factors considered for inclusion. This heuristic and guideline consisted of two components: (1.a) provide the

policy before entry and/or suggestions during entry, and (1.b) keep them visible during the password creation process. Overall, almost all users (95.83%) encountered instances of usability problems (13.30%) with the provision and timing of presentation in all PCSs, and the mean severity rating of these problems was high (2.74).

Table 10.2 Components of the first heuristic and guideline and rationale for inclusion

Heuristic/ Guideline	Code Component	#1	1.a	1.b
			Provide policy before entry and/or suggestions during entry	Keep them visible during password creation process
<i>User problem category</i>				
			1.1 Not provided or provided too late 2.1 Not provided or provided too late	2.6 Complex presentation 6.2 Poor/unclear/confusing presentation
	<i>No (%) of users (N=24)</i>	23 (95.83)	22 (91.67)	12 (50)
	<i>No (%) of PCSs (N=6)</i>	6 (100)	6 (100)	3 (50)
<i>No (%) of problems</i>				
	Distinct (N=81)	13 (16.05)	10 (12.35)	3 (3.70)
	Instances (N=654)	87 (13.30)	72 (11.01)	15 (2.29)
	<i>Mean severity rating (SD)</i>	2.74 (0.96)	2.84 (0.93)	2.31(1.03)
<i>Example comments</i>				
			<i>It doesn't give any requirements (P05)</i> <i>I expect the system to help me to improve my password. (P14)</i>	<i>The dialog box needs to always be there and not disappear when I move to another field. (P03)</i>

Note. The data in this table based on data from Study 3, Table 5.4 in Chapter 5.

The findings from users' perceptions about current PCS practises further support the first heuristic and guideline, as shown in Table 10.3. Overall, 20 best practices were reported about providing the statements of policy and creation suggestion in PCSs, and 15 worst practices were reported about not providing them.

The first heuristic and guideline were also supported by evidence from the users' performance in terms of the best timing of presentation. For the password policy, the findings revealed that presentation at the before-interaction step had the following effects on users and the passwords they created: users required less time and effort to create passwords and had high successful recall rates of their passwords; and passwords were long (not as long as those created at the after-interaction step, but on average those created at the before-interaction step were one character shorter),

contained all the character classes, and 70.59% were compliant. For the password creation suggestion, the results showed that presentation at the during-interaction step impacted the users and the passwords they created in similar ways to the policy: less time was required with high levels of perceived confidence; and passwords were long and contained full character classes.

Table 10.3 Supporting evidence from the users' perception data regarding the first heuristic and guideline

		No (%) of users (N=24)	No (%) of PCSs (N=6)	No (%) of best practices	No (%) of worst practices	Example comments
Appreciation for these supporting features	<i>policy</i>	10 (41.67)	4 (66.67)	15/17 (88.24)	-	<i>They tell you what you need from the beginning. (P23)</i>
	<i>creation suggestion</i>	4 (16.67)	2 (33.33)	5/10 (50)	-	<i>I like the handy tips at the side very clear. (P12)</i>
Disappointment with lack of supporting features	<i>policy</i>	9 (37.50)	4 (66.67)	-	13/29 (44.83)	<i>It doesn't give you any guidelines on how to make a password. (P02)</i>
	<i>creation suggestion</i>	1 (4.17)	2 (33.33)	-	2/8 (25)	<i>It doesn't give you advice or help you. (P12)</i>

Note. The data in this table based on data from Study 3, Table 5.5 and Table 5.6 in Chapter 5.

Heuristic and Guideline #2

Provide a clear, sufficiently detailed and logical statement of the password policy and password creation suggestions. Provide the same for error messages.

- The password policy and creation suggestions should be stated in specific, clear, and easy-to-understand terms. Provide enough detail without overwhelming the user with too much information. The same principles apply to error messages associated with password creation.
- Make sure that the logic of statements, such as what constitutes a valid and invalid password, is completely clear.
- Be consistent in terminology and avoid ambiguous terms or jargon.
- The password policy should be given in a declarative form before password entry (e.g. 'The password needs to include at least six characters'), but in a procedural form during and after password entry (e.g. 'Include at least six characters').
- Password creation suggestions and error messages should be in declarative forms throughout the password creation process (e.g. creation suggestion: 'You can improve your password by having a combination of numbers, letters and symbols like!@#~').

Table 10.4 Components of the second heuristic and guideline and rationale for inclusion

Heuristic/ Guideline	Code Component	#2	2.a	2.b	2.c	2.d	2.e	2.f
			Clear/easy to understand	Enough detail, not too much	Clear logic	Consistent terminology	Avoid jargon	Be specific
<i>User problem category</i>			1.4 Confusing/odd statement	1.2 Information not detailed enough 1.3 Information too detailed 2.2 Information not detailed enough 2.3 Information too detailed	1.4 Confusing/ odd statement 5.3 Inconsistent statement	1.5 Inconsistent terms 5.3 Inconsistent statement	1.4 Confusing/ odd statement 5.6 Unclear/ confusing language	5.1 Information not specific enough
<i>No (%) of users (N=24)</i>		24 (100)	22 (91.67)	22 (91.67)	14 (58.22)	19 (79.17)	16 (66.67)	19 (79.17)
<i>No (%) of PCS (N=6)</i>		5 (83.33)	2 (33.33)	5 (83.33)	2 (33.33)	2 (33.33)	2 (33.33)	2 (33.33)
<i>No (%) of problems</i>								
Distinct (N=81)		24 (29.63)	7 (8.64)	8 (9.88)	2 (2.47)	2 (2.47)	2 (2.47)	3 (3.70)
Instances (N=654)		203 (31.04)	51 (7.80)	64 (8.79)	16 (2.45)	22 (3.36)	22 (3.36)	28 (4.28)
<i>Mean severity rating (SD)</i>		2.67 (0.87)	2.62 (0.92)	2.64 (0.88)	2.47 (0.99)	2.57 (1.03)	2.91 (0.75)	2.71 (0.75)
<i>Example comments</i>			<i>It's not clear. I added what it tells me to add but my password still has a problem.</i> (P11)	<i>It seems too specific and it's asking a lot of me. I'll probably read the first three and ignore the rest of it.</i> (P17)	<i>It asks you for two of them not all of them. What happened to the special characters when I added uppercase and lowercase?</i> (P16)	<i>They should be consistent in using the words symbols and special characters.</i> (P14)	<i>The word 'additional' is a bit vague. It doesn't specify which character I should add.</i> (P17)	<i>Of this is the same message - this is terrible it's a general message, not specific to the problem with my password, so it's not helpful.</i> (P03)

Note. The data in this table based on data from Study 3, Table 5.4 in Chapter 5.

The second heuristic and guideline concerned the clarity and phrasing of the policy, creation suggestions, and error message statements. It was developed mainly from the usability problems users encountered in Study 3 and supported by evidence from user perception in Studies 3 and 5. Table 10.4 illustrates the components of the second heuristic and guideline along with the factors considered for inclusion.

Table 10.5 Supporting evidence from the users' perception data regarding the second heuristic and guideline

		No (%) of users (N=24)	No (%) of PCSs (N=6)	No (%) of best practices	No (%) of worst practices	Example comments
Appreciation for how these statements are stated	<i>policy</i>	8 (33.33)	3 (50)	8/17 (47.06)	-	<i>The list here is so specific.</i> (P04)
	<i>creation suggestion</i>	4 (16.67)	2 (33.33)	4/10 (40)	-	<i>The use of suggest at the beginning. It's quite casual and was not complicated.</i> (P17)
	<i>error message</i>	9 (37.50)	2 (33.33)	10/14 (71.43)	-	<i>The fact that it generates a specific response to what I just typed.</i> (P12)
Disappointment with how these statements are stated	<i>policy</i>	9 (37.50)	4 (66.67)	-	14/29 (48.28)	<i>The lack of definite instruction. Everything seems to be a suggestion.</i> (P18)
	<i>creation suggestion</i>	1 (4.17)	2 (33.33)	-	3/8 (37.50)	<i>Clarity on what makes a strong password.</i> (P16)
	<i>error message</i>	10 (41.67)	4 (66.67)	-	13/18 (72.22)	<i>When I get it wrong it doesn't tell me what I've done wrong.</i> (P03)

Note. The data in this table based on data from Study 3, Table 5.5 and Table 5.6 in Chapter 5.

This heuristic and guideline consisted of six components: (2.a) clear and easy to understand, (2.b) enough detail, not too much, (2.c) clear logic, (2.d) consistent terminology, (2.e) avoid jargon, and (2.f) be specific. Overall 31.04% of instances of usability problems related to the clarity and phrasing of the statements of policy and creation suggestions. All users reported problems in this area. These usability problems occurred in almost all PCSs (83.33%) and had a high mean severity rating (2.67).

The users' perceptions about the current PCS practices further support the second heuristic and guideline, as shown in Table 10.5. Overall, 22 best practices were reported about the phrasing of the statements of policy, creation suggestions, and error messages in PCSs, while 30 were reported about the worst practices. Furthermore, the second

heuristic and guideline were supported by evidence from Study 5, which concerned the format used for phrasing the policy, creation suggestion, and error message statements across the different timings of presentation. In terms of the policy statements, the findings revealed that users perceived the declarative format at the before-interaction step to be clearer and more helpful, and they also felt more confident reading this format. The same was true for the procedural format at the during-interaction and after-interaction steps. For the suggestion and error message statements, users gave high ratings on helpfulness, clarity, amount of detail, and their confidence when reading the declarative format at the during-interaction and after-interaction steps.

Heuristic and Guideline #3

Provide basic information about required password composition in the policy.

- The system should provide information regarding the minimum length of valid passwords (and the maximum length, if applicable) and the required character classes (e.g. numbers, capital letters, symbols).

The third heuristic and guideline concerned the provision of information about password composition and requirements. It was developed mainly from the usability problems users encountered in Study 3, and supported by other evidence from the same study. Table 10.6 illustrates the factors considered for the inclusion of this heuristic and guideline.

Table 10.6 The third heuristic and guideline and rationale for inclusion

Heuristic/ Guideline	Code Component	#3 -
<i>User problem category</i>		1.1 Not provided or provided too late
<i>No (%) of users (N=24)</i>		17 (70.83)
<i>No (%) of PCSs (N=6)</i>		2 (33.33)
<i>No (%) of problems</i>		
	Distinct (N=81)	2 (2.47)
	Instances (N=654)	26 (3.98)
<i>Mean severity rating (SD)</i>		2.69 (0.98)
<i>Example comments</i>		<i>It would be handy if it had this to start with instead of guessing how many characters they want. (P18)</i>

Note. The data in this table based on data from Study 3, Table 5.4 in Chapter 5.

Overall, most users (70.83%) encountered usability problems (3.98% of all instances) with password requirements. These problems occurred in two PCSs (33.33%) and had a high mean severity rating (2.69). Furthermore, as shown in Table 10.7, the users' perceptions about the current practices of PCSs (from Study 3) also support the third heuristic and guidelines. Only two best practices were reported about informing the users of what was needed in the passwords, while four worst practices were reported.

Table 10.7 Supporting evidence from the users' perception data regarding the third heuristic and guideline

		No (%) of users (N=24)	No (%) of PCSs (N=6)	No (%) of best practices	No (%) of worst practices	Example comments
Appreciation for information about requirement	<i>policy</i>	2 (8.33)	2 (33.33)	2/17 (11.76)	-	<i>The best thing that they tell me is how long it needs to be. (P24)</i>
Disappointment with lack of information about requirement	<i>policy</i>	4 (16.67)	2 (33.33)	-	4/29 (13.79)	<i>There is nothing about the password complexity except the length. (P12)</i>

Note. The data in this table based on data from Study 3, Table 5.5 and Table 5.6 in Chapter 5.

Heuristic and Guideline #4

Make careful use of dynamic presentation of the password policy and password creation suggestions.

- When using a dynamic presentation of the password policy, make sure this will be easily understood by users and will not distract them. An example of dynamic presentation is when elements required by the policy are displayed before the user starts entering her password, and that display then changes as elements are included in the password.
- Do not remove elements required by the policy from the display when the user's password includes that required element.
- Do not use colour only to indicate that an element has been included. Good practice is to change colour (e.g. from red to green) **and** indicate that the element has been included with a tick.

The fourth heuristic and guideline concerned the use of dynamic presentation of the policy and creation suggestion statements. This heuristic was developed mainly from the usability problems users encountered in Study 3 and supported by other evidence

from the same study. Table 10.8 illustrates the components of the fourth heuristic and guideline and the factors considered for inclusion.

Table 10.8 Components of the fourth heuristic and guideline and rationale for inclusion

Heuristic/ Guideline	Code Component	#4	4.a	4.b
			Make it noticeable	Make it legible
<i>User problem category</i>			1.4 Confusing/odd statement 1.7 Poor/unclear/confusing presentation 2.7 Poor/unclear/confusing presentation 6.2 Poor/unclear/confusing presentation	6.2 Poor/unclear/confusing presentation
<i>No (%) of users (N=24)</i>		18 (75.00)	16 (66.67)	5 (20.83)
<i>No (%) of PCSs (N=6)</i>		4 (66.67)	3 (50)	2 (33.33)
<i>No (%) of problems</i>				
	Distinct (N=81)	7 (8.64)	5 (6.17)	2 (2.47)
	Instances (N=654)	28 (4.28)	23 (3.52)	5 (0.76)
<i>Mean severity rating (SD)</i>		2.58 (1.02)	2.58 (1.02)	2.60 (1.14)
<i>Example comments</i>			<i>I didn't notice that there's a clickable link. I want everyone to see it and it should be highlighted more. (P23)</i>	<i>The field is very small and the name of the field disappears when I type in it. (P22)</i>

Note. The data in this table based on data from Study 3, Table 5.4 in Chapter 5.

Table 10.9 Supporting evidence from the users' perception data regarding the fourth heuristic and guideline

		No (%) of users (N=24)	No (%) of PCSs (N=6)	No (%) of best practices	Example comments
Appreciation for the use of dynamic presentation	<i>Other feedback</i>	18 (75)	2 (33.33)	26/40 (65)	<i>Having the tick boxes with the green light that showed up as you fulfilled the criteria. (P08)</i>

Note. The data in this table based on data from Study 3, Table 5.5 in Chapter 5.

This heuristic and guideline consisted of two components: (4.a) make it noticeable and (4.b) make it legible. Overall, most users (75.00%) encountered usability problems (4.28% of all instances) in relation to the dynamic presentation of the statement policy and creation suggestions. These problems occurred in four PCSs (66.67%) and had a high mean severity rating (2.58). The users' perceptions about the current PCS practices

(from Study 3) further support the fourth heuristic and guideline, as shown in Table 10.9. Twenty-six best practices were reported about the use of dynamic presentation.

<i>Heuristic and Guideline #5</i>	
If possible, provide information about the strength of the password.	
<ul style="list-style-type: none"> • It is helpful if the system provides information about the strength of the password (e.g. a password meter), why the password has been assigned this level of strength, and how to make it stronger. • Strength should be indicated both graphically (e.g. with a three-colour, traffic light metaphor) and in text. 	

The fifth heuristic and guideline concerned the provision of a password strength indicator and its presentation. This heuristic and guideline was developed mainly from the usability problems users encountered in Study 3 and supported by evidence from user perception (Study 3) and user performance (Study 6). Table 10.10 illustrates the components of the fifth heuristic and guideline and the factors considered for inclusion.

Table 10.10 Components of the fifth heuristic and guideline and rationale for inclusion

Heuristic/ Guideline	Code Component	#5	5.a	5.b	5.c
			Indicate strength	Specify why password has this strength	Provide ways to improve the password
<i>User problem category</i>			3.1 Not provided	6.2 Poor/ unclear/ confusing presentation	2.4 Unclear how to create a good password
	<i>No (%) of users (N=24)</i>	24 (100)	13 (54.17)	22 (91.67)	6 (25)
	<i>No (%) of PCSs (N=6)</i>	6 (100)	4 (66.67)	1 (16.67)	1 (16.67)
	<i>No (%) of problems</i>				
	Distinct (N=81)	7 (8.64)	4 (4.94)	2 (2.47)	1 (1.23)
	Instances (N=654)	61 (9.33)	20 (3.06)	35 (5.35)	6 (0.92)
	<i>Mean severity rating (SD)</i>	2.99 (0.98)	3.00 (1.12)	3.05 (0.91)	2.50 (1.00)
	<i>Example comments</i>		<i>It doesn't give me any sort of scale to gauge whether or not it was a secure password (P22)</i>	<i>I feel this password is strong, but why is it weak? (P03)</i>	<i>It's confusing. It gives me an indicator of medium and most of the dialogue is ticked. (P11)</i>

Note. The data in this table based on data from Study 3, Table 5.4 in Chapter 5.

This heuristic and guideline consisted of three components: (5.a) provide the strength, (5.b) specify why the password is given this strength, and (5.c) provide a way to improve the password. Overall, all users encountered usability problem instances (9.33%) with the password strength indicator; these occurred in all six PCSs and had a high mean severity rating (2.99). Furthermore, as shown in Table 10.11, the users' perceptions about the current PCS practices (Study 3) also support the fifth heuristic and guideline. The respondents reported five best practices about providing a strength indicator, and 13 worst practices concerning not providing it or how it was provided.

Table 10.11 Supporting evidence from the users' perception data regarding the fifth heuristic and guideline

		No (%) of users (N=24)	No (%) of PCSs (N=6)	No (%) of best practices	No (%) of worst practices	Example comments
Appreciation for strength indicator	<i>Strength indicator</i>	5 (20.83)	2 (33.33)	5/6 (83.33)	-	<i>Probably the indicators [meter] (traffic light) when it comes up when you're creating a password. (P10)</i>
Disappointment with lack of strength indicator	<i>Strength indicator</i>	8 (33.33)	5 (83.33)	-	13/14 (91.67)	<i>I quite like to see a visual thing that tells me how strong my password is. (P12)</i>

Note. The data in this table based on data from Study 3, Table 5.5 and Table 5.6 in Chapter 5.

The fifth heuristic and guideline were also supported by evidence from the experimental user study (Study 6) in terms of the colour-scheme and media presentation of the password strength indicator. For the colour-scheme, presenting a *3colour* indicator impacted the user satisfaction and passwords as follows: a higher level of perceived clarity was found; and passwords were long and contained a high number of digits with a high strength score. Within the *3colour* indicator, the user study showed that presenting a *graphical&textual* indicator had the following effects on users and passwords: users felt highly confident using this indicator, rated the perceived clarity as high, and had high successful recall rates; and passwords were long and contained a high number of digits and symbols with a high strength score.

*Heuristic and Guideline #6***Provide a clear indication of whether a password is valid or not.**

- Clearly differentiate between an invalid password and a weak but valid password.

The sixth heuristic and guideline concerned the acknowledgement of the password validity and clarity in giving feedback, and the presentation of this feedback. This heuristic and guideline was developed mainly from the usability problems users encountered and supported by evidence from Study 3. Table 10.12 illustrates the components of the sixth heuristic and guideline and the factors considered for inclusion.

Table 10.12 Components of the sixth heuristic and guideline and rationale for inclusion

Heuristic/ Guideline	Code Component	#6	6.a	6.b
			Provide feedback on the validity of the password	Be clear on the validity of the password and its strength level
<i>User problem category</i>				
			4.2 No feedback about valid/invalid password	3.6 Querying the strength indicators 5.2 Poor/unclear/confusing presentation
	<i>No (%) of users (N=24)</i>	24 (100)	21 (87.50)	23 (95.83)
	<i>No (%) of PCSs (N=6)</i>	5 (83.33)	4 (66.70)	2 (33.33)
	<i>No (%) of problems</i>			
	Distinct (N=81)	6 (7.41)	4 (4.94)	2 (2.47)
	Instances (N=654)	92 (14.07)	45 (6.88)	47 (7.19)
	<i>Mean severity rating (SD)</i>	2.91 (0.87)	2.90 (0.96)	2.93 (0.76)
	<i>Example comments</i>			
			<i>It's hard to tell and confusing whether it lets you have this password or not. (P01)</i>	<i>It tricks me and is just confusing. Strong beside a red exclamation mark - having two contradictory things is not too clear. (P18)</i>

Note. The data in this table based on data from Study 3, Table 5.4 in Chapter 5.

This heuristic and guideline consisted of two components: (6.a) provide feedback on the validity of the password and (6.b) be clear on the validity of the password and its strength level. Overall, all users encountered usability problem instances (14.07%) with the validity feedback. These problems occurred in five PCSs (83.33%) and had a high mean severity rating (2.91). As shown in Table 10.13, the users' perception about the

current PCS practices further support the sixth heuristic and guideline. Thirteen best practices were reported about providing password validity feedback, while sixteen worst practices were reported of not providing any feedback.

Table 10.13 Supporting evidence from the users' perception data regarding the sixth heuristic and guideline

		No (%) of users (N=24)	No (%) of PCSs (N=6)	No (%) of best practices	No (%) of worst practices	Example comments
Appreciation for validity feedback	<i>Other feedback</i>	9 (37.50)	3 (50)	13/40 (32.50)	-	<i>It checks the password and if it's okay you get a tick. And if it's not okay you get an explanation underneath. (P11)</i>
Disappointm ent with lack of validity feedback	<i>Other feedback</i>	16 (66.67)	5 (83.33)	-	28/66 (42.42)	<i>When it's telling me 'you can't use it' and still telling me it's strong or medium. (P01)</i>

Note. The data in this table based on data from Study 3, Table 5.5 and Table 5.6 in Chapter 5.

Heuristic and Guideline #7

Include password re-entry and positive confirmation.

- Ask the user to re-enter his password for validation. Provide positive confirmation that the two entries match, as well as feedback on the lack of a match.

The seventh heuristic and guideline concerned the provision of an entry field to confirm the proposed password. This heuristic and guideline was developed mainly from the usability problems users encountered in Study 3 and supported by evidence from the same study. Table 10.14 illustrates the components of the seventh heuristic and guideline and the factors considered for inclusion. This heuristic and guideline consisted of two components: (7.a) provide confirmation field and (7.b) provide feedback on matching/non-matching passwords. Overall, almost all users (95.83%) encountered usability problems (8.26% of all instances) with the provision of the password confirmation field. These problems occurred in five PCSs (83.33%) and had a high mean severity rating (2.81).

Table 10.14 Components of the seventh heuristic and guideline and rationale for inclusion

Heuristic/ Guideline	Code Component	#7	7.a	7.b
			Provide confirmation field	Provide feedback on matching/non-matching passwords
<i>User problem category</i>			6.6 Functionally lacking	4.3 No feedback about matching/non-matching password
<i>No (%) of users (N=24)</i>		23 (95.83)	22 (91.67)	14 (58.33)
<i>No (%) of PCSs (N=6)</i>		5 (83.33)	2 (33.33)	3 (50)
<i>No (%) of problems</i>				
Distinct (N=81)		5 (6.17)	2 (2.47)	3 (3.70)
Instances (N=654)		54 (8.26)	31 (4.74)	23 (3.52)
<i>Mean severity rating (SD)</i>		2.81 (0.84)	2.97 (0.68)	2.58 (1.02)
<i>Example comments</i>			<i>Not having a confirm password; there's a big risk of making a typo and getting it wrong. (P18)</i>	<i>Is this confirmed? Is that okay? No indication. (P15)</i>

Note. The data in this table based on data from Study 3, Table 5.4 in Chapter 5.

Table 10.15 Supporting evidence from the users' perception data regarding the seventh heuristic and guideline

		No (%) of users (N=24)	No (%) of PCSs (N=6)	No (%) of best practices	No (%) of worst practices	Example comments
Appreciation for confirmation field	<i>Other comments</i>	3 (12.50)	2 (33.33)	4/70 (5.71)	-	<i>Has a second confirm password box. (P20)</i>
Disappointment with lack of confirmation field	<i>Other comments</i>	3 (12.50)	2 (33.33)	-	3/28 (10.71)	<i>I don't like the lack of confirm password field. (P09)</i>
	<i>Other feedback</i>	2 (8.33)	1 (16.67)	-	2/66 (3.03)	<i>It doesn't say whether it's a match or not. (P06)</i>

Note. The data in this table based on data from Study 3, Table 5.5 and Table 5.6 in Chapter 5.

As shown in Table 10.15, the users' perception about the current PCS practices further support the seventh heuristic and guideline. The respondents reported only four best practices about providing a password confirmation field, and five worst practices of not providing one or not providing positive/negative match feedback.

Heuristic and Guideline #8

Provide at least two of the following features to support the user in the password creation process: password policy, password creation suggestions, and an indication of password strength (e.g. a password meter).

- The system should provide at least two of the following features: password policy, password creation suggestions, and an indication of password strength (e.g. a password meter).

The eighth heuristic and guideline were derived from the comparison of the individual and combined effects of the supporting features (Studies 6 and 7). The perceived usability of the PCSs significantly improved when the password policy was combined with a strength indicator or creation suggestions, more than when the password policy was presented alone. It does not seem to matter what supporting features are presented together, as long as they are not presented alone.

10.2.3 Review Process of the First Version of the *PassHeuristics* and *PassGuidelines*

Figure 10.2 illustrates the process of reviewing several versions of the *PassHeuristics* and *PassGuidelines* until a final version emerged.

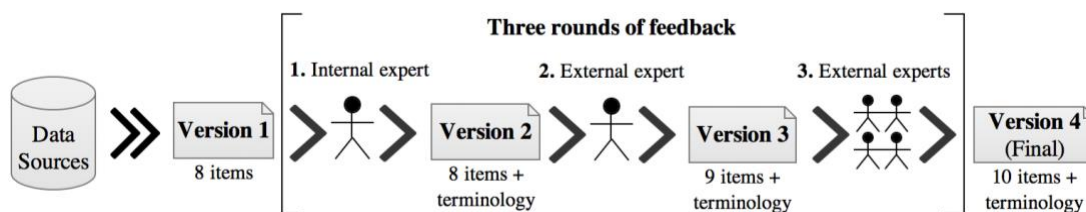


Figure 10.2 Process of reviewing the versions of the *PassHeuristics* and *PassGuidelines*

The first version went through three rounds of thorough feedback with a total of six usability professionals. The feedback concerned the clarity and understandability of the proposed set of heuristics and guidelines. The first round of feedback was performed by an internal usability expert who was involved in this research programme, and the second by an external usability expert who was not involved in the research programme. Finally, four usability experts performed the third round via an online survey.

All six usability experts worked or studied at the University of York. There were three women and three men, and their ages ranged from 23 to 66 with an average of 48.50 years (standard deviation = 15.88). All experts had a degree relevant to usability and had at least four years' experience. The experts' participation was voluntary and they were not compensated for it.

10.2.3.1 First Round of Feedback

In the first round, the internal expert suggested adding a terminology list before presenting the full set of heuristics and guidelines to improve clarity and introduce the terms used in it to users such as evaluators, developers, and/or designers. Therefore, the second version included a list of terms and their definitions. The terminology list comprised the following: password creation systems, password policy, password creation suggestion, and password strength indicator. Table 10.16 presents the second version of the *PassHeuristics* and *PassGuidelines* developed after the first round of feedback.

10.2.3.2 Second Round of Feedback

In the second round of feedback, the external expert found the *PassHeuristics* and *PassGuidelines* difficult to understand. There were two concerns with this second version. The first concern regarded the terms used in the heuristics and guidelines.

The two terms 'password creation system' and 'password creation suggestion' were found to be very similar and therefore easily confused. Therefore, in the third version, the term 'password creation system' was shortened and referred to only as 'system'. The second concern was about the first and last heuristics and guideline: the expert found them contradictory. The first heuristic and guideline in the second version covered the effective timing of presentation for both the statement of policy and creation suggestions. It indicated that the policy should be provided before password entry, and suggestions during password entry, and that both should be kept visible throughout the password creation process. On the other hand, the last heuristic and

guideline in the second version concerned providing at least two of the following features in a PCS: policy, creation suggestions, and a password strength indicator. The expert found this to be contradictory, as the first heuristic and guideline suggested the necessity of providing both policy and suggestions as they were covered together, while the last one indicated that any combination could be provided.

Therefore, in the third version, two changes were made to address this concern. The first change was dividing the first heuristic and guideline in the second version into two separate heuristics and guidelines. One heuristic and guideline covered the timing of presentation for the statement of policy only, and the other the statement of creation suggestions only. Furthermore, attention was paid to the language used to phrase the two heuristics and guidelines; thus, the verbs ‘should’ and ‘may’ were used for the policy and creation suggestions, respectively. The second change was ordering the heuristic and guideline items: the last heuristic and guideline became the first one, which should give users information about the supporting features and their possible combinations in PCSs before reading the rest of the items. Finally, the third version of the *PassHeuristics* and *PassGuidelines* contained nine items. Table 10.17 presents the third version of the heuristics and guidelines emerged after the second round of feedback.

Table 10.16 Second version of the *PassHeuristics* and *PassGuidelines* after the first round of feedback

<p style="text-align: center;"><i>PassHeuristics</i> and <i>PassGuidelines</i> (Version 2):</p> <p style="text-align: center;">Heuristics and Guidelines for the Evaluation and Development of Password Creation Systems</p> <p>Terminology used in these heuristics and guidelines:</p> <ul style="list-style-type: none"> ○ <i>Password creation system</i>: interactive system which allows users to create a password ○ <i>Password policy</i>: a set of rules that determine the content of valid passwords for a given system ○ <i>Password creation suggestion</i>: advice to users on the content and structure of good passwords ○ <i>Password strength indicator</i>: a visualization of an estimate of the strength of a proposed password (e.g. a password meter). <p>#1. Provide the password policy before the users start entering their password and leave it available to them during password entry and after entry is completed. Provide password creation suggestions only once password entry has begun and leave it available during password entry and after entry is completed.</p> <ul style="list-style-type: none"> • The system should always provide the user with a statement of the password policy as soon as they land on the page. • The system may also provide password creation suggestions as soon as the user starts typing their password. • In both cases, the information should remain visible throughout the password creation process. <p>#2. Provide a clear, sufficiently detailed and logical statement of the password policy and password creation suggestions. Provide the same for error messages.</p> <ul style="list-style-type: none"> • The password policy and creation suggestions should be stated in specific, clear and easy to understand terms. Provide enough detail without overwhelming the user with too much information. The same principles apply to error messages associated with password creation. • Make sure that the logic of statements, such as what constitutes a valid and invalid password, is completely clear. • Be consistent in terminology and avoid ambiguous or jargon terms. • The password policy should be given in a declarative form before password entry (e.g. The password needs to include at least six characters), but in procedural form during and after password entry (e.g. Include at least six characters). • Password creation suggestions and error messages should be in declarative form throughout the password creation process (e.g. creation suggestion: You can improve your password by having a combination of numbers, letters and symbols like !@#~) 	<p>#3. Provide basic information about required password composition in the policy</p> <ul style="list-style-type: none"> • The system should provide information regarding the minimum length of valid passwords (and the maximum length, if applicable) and the required character classes (e.g. numbers, capital letters, symbols). <p>#4. Make careful use of dynamic presentation of the password policy and password creation suggestions</p> <ul style="list-style-type: none"> • If you decide to use dynamic presentation of the password policy, make sure this will be easily understood by users and not distract them. An example of dynamic presentation is when elements required by the policy are displayed before the user starts entering their password and as elements are included in the password, the display of the policy changes. • Do not remove elements required by the policy from the display when the user's password includes that required element. • Do not use colour only to indicate that an element has been included. Good practice is to change colour (e.g. from red to green) and indicate the element has been included with a tick. <p>#5. If possible provide information about the strength of the password</p> <ul style="list-style-type: none"> • It is helpful if the system can provide information about the strength of the password (e.g. a password meter), why the password is assigned this level of strength and how to make it stronger. • Strength should be indicated both graphically (e.g. with a three colour, traffic light metaphor) and in text. <p>#6. Provide clear indication of whether a password is valid or not</p> <ul style="list-style-type: none"> • Also differentiate clearly between a valid password and a weak but valid password. <p>#7. Include password re-entry and positive confirmation</p> <ul style="list-style-type: none"> • Ask the user to re-enter their password for validation. Provide positive confirmation that the two entries match, as well as feedback on lack of match. <p>#8. Provide at least two of the following features to support the user in password creation: password policy, password creation suggestions, an indication of password strength (e.g. a password meter).</p>
---	--

Table 10.17 Third version of the *PassHeuristics* and *PassGuidelines* after the second round of feedback

<i>PassHeuristics and PassGuidelines (Version 3):</i>	
<p style="text-align: center;">Heuristics and Guidelines for the Evaluation and Development of Password Creation Systems</p> <p>Terminology used in these heuristics and guidelines:</p> <ul style="list-style-type: none"> ○ <i>Password creation system</i>: interactive system which allows users to create a password, this will be referred to as the system ○ <i>Password policy</i>: a set of rules that determine the content of valid passwords for a given system ○ <i>Password creation suggestion</i>: advice to users on the content and structure of good passwords ○ <i>Password strength indicator</i>: a visualization of an estimate of the strength of a proposed password (e.g. a password meter). <p>#1. Provide at least two of the following features to support the user in password creation: password policy, password creation suggestions, an indication of password strength (e.g. a password meter).</p> <p>#2. Provide the password policy before the users start entering their password and leave it available to them during password entry and after entry is completed.</p> <ul style="list-style-type: none"> • The system should always provide the user with a statement of the password policy as soon as they open the system, and the information should remain visible throughout the password creation process. <p>#3. Provide password creation suggestions only once password entry has begun and leave it available during password entry and after entry is completed.</p> <ul style="list-style-type: none"> • The system may provide password creation suggestions as soon as the user starts typing in their password, and the information should remain visible throughout the password creation process. <p>#4. Provide a clear, sufficiently detailed and logical statement of password policy and password creation suggestions. Provide the same for error messages.</p> <ul style="list-style-type: none"> • The password policy and creation suggestions should be stated in specific, clear and easy to understand terms. Provide enough detail without overwhelming the user with too much information. The same principles apply to error messages associated with password creation. • Make sure that the logic of statements, such as what constitutes a valid and invalid password, is completely clear. • Be consistent in terminology and avoid ambiguous or jargon terms. • The password policy should be given in a declarative form before password entry (e.g. The password needs to include at least six characters), but in procedural form during and after password entry (e.g. Include at least six characters). • Password creation suggestions and error messages should be in declarative form throughout the password creation process (e.g. You can improve your password by having a combination of numbers, letters and symbols like !@#~) 	<p>#5. Provide basic information about required password composition in the policy</p> <ul style="list-style-type: none"> • The system should provide information regarding the minimum length of valid passwords (and the maximum length, if applicable) and the required character classes (e.g. numbers, capital letters, symbols). <p>#6. Make careful use of dynamic presentation of the password policy and password creation suggestions</p> <ul style="list-style-type: none"> • If you decide to use dynamic presentation of the password policy, make sure this will be easily understood by users and not distract them. An example of dynamic presentation is when elements required by the policy are displayed before the user starts entering their password and as elements are included in the password, the display of the policy changes. • Do not remove elements required by the policy from the display when the user's password includes that required element. • Do not use colour only to indicate that an element has been included. Good practice is to change colour (e.g. from red to green) and indicate the element has been included with a tick. <p>#7. If possible provide information about the strength of the password</p> <ul style="list-style-type: none"> • It is helpful if the system can provide information about the strength of the password (e.g. a password meter), why the password is assigned this level of strength and how to make it stronger. • Strength should be indicated both graphically (e.g. with a three colour, traffic light metaphor) and in text. <p>#8. Provide clear indication of whether a password is valid or not</p> <ul style="list-style-type: none"> • Also differentiate clearly between an invalid password and a weak but valid password. <p>#9. Include password re-entry and positive confirmation</p> <ul style="list-style-type: none"> • Ask the user to re-enter their password for validation. Provide positive confirmation that the two entries match, as well as feedback on lack of match.

10.2.3.3 Third Round of Feedback

As the third round of feedback, an online survey was conducted with four external usability experts on the third version of the proposed *PassHeuristics* and *PassGuidelines*. Experts were asked to read through the heuristics and guidelines and comment on each individual item and the overall set.

The survey collected information about the experts' opinion of the individual item by measuring the following dependent variables: (1) perceived ease of understanding, (2) perceived clarity, (3) perceived amount of detail, (4) perceived ease of use, (5) perceived usefulness, and (6) their confidence in evaluating a PCS using this heuristic. The survey also collected information about the experts' overall opinion of the proposed heuristics and guidelines by measuring (7) the intention to use them and (8) the perceived completeness. Each measure in the survey was formulated using a 5-point Likert item (higher = better). The dependent measures (4), (5), (7), and (8) have been used in the literature for heuristic evaluation (Paz, Paz, & Pow-Sang, 2015). Table 10.18 summarises the results for the dependent measures for each item.

Table 10.18 Mean (median) ratings of the six dependent measures for individual items of the *PassHeuristics* and *PassGuidelines*

	Ease of understanding	Clarity	Amount of detail	Ease of use	Usefulness	Confidence
#1	4.25 (4.00)	4.00 (4.00)	2.75 (3.00)	4.00 (4.00)	3.75 (4.00)	4.50 (4.50)
#2	3.00 (2.50)	3.75 (3.50)	3.75 (3.50)	3.25 (3.50)	3.50 (3.50)	4.00 (4.00)
#3	3.75 (4.00)	3.50 (3.50)	3.00 (3.00)	3.25 (3.50)	3.25 (3.50)	3.50 (3.50)
#4	3.50 (3.50)	3.75 (4.00)	4.25 (4.00)	2.75 (3.00)	3.50 (4.00)	2.75 (2.50)
#5	3.75 (4.00)	4.00 (4.00)	2.50 (2.50)	4.00 (4.00)	4.00 (4.00)	3.75 (4.00)
#6	4.00 (4.00)	3.75 (4.00)	3.50 (3.50)	3.75 (4.00)	4.00 (4.00)	3.75 (3.50)
#7	4.00 (4.00)	4.25 (4.50)	3.00 (3.00)	4.00 (4.00)	3.75 (4.50)	3.75 (3.50)
#8	4.25 (4.00)	4.50 (4.50)	2.25 (2.50)	3.75 (4.50)	3.25 (3.50)	4.25 (5.00)
#9	4.75 (5.00)	4.75 (5.00)	3.00 (3.00)	5.00 (5.00)	5.00 (5.00)	4.75 (5.00)
<i>p value</i>	n.s.	n.s.	.007	n.s.	n.s.	n.s.
<i>Overall</i>	3.92 (3.88)	4.03 (4.56)	3.11 (3.11)	3.75 (3.94)	3.78 (4.00)	3.89 (3.94)

Friedman's test was used for each measure to compare the ratings between the nine items of the *PassHeuristics* and *PassGuidelines*. The results showed a significant difference only in the ratings of the amount of detail ($\chi^2(8) = 21.58$, $p = .006$) between the individual heuristics and guidelines. However, no significant difference was found in the ratings of ease of understanding ($\chi^2(8) = 6.77$, $p = .562$), clarity ($\chi^2(8) = 6.59$, $p = .581$), ease of use ($\chi^2(8) = 10.34$, $p = .242$), usefulness ($\chi^2(8) = 9.07$, $p = .336$), or confidence ($\chi^2(8) = 8.70$, $p = .368$). A one-sample Wilcoxon Signed Rank test was used for each measure to compare the observed median against the hypothesised median (midpoint = 3) across the individual item of the heuristics and guidelines, and the overall *PassHeuristics* and *PassGuidelines*. The results indicated the following:

- Clarity: the rating of clarity was significantly higher than the midpoint for the first heuristic and guideline ($Z = 10.00$, $p = .046$).
- Ease of use: the rating of ease of use was significantly higher than the midpoint for the ninth heuristic and guideline ($Z = 10.00$, $p = .046$).
- Usefulness: the rating of usefulness was significantly higher than the midpoint for the ninth heuristic and guideline ($Z = 10.00$, $p = .046$).

Overall, the experts rated their likelihood of using the proposed heuristics and guidelines in the future as very likely ($M = 4.00$, $Mdn = 4.50$). Furthermore, the *PassHeuristics* and *PassGuidelines* were perceived to be complete ($M = 3.50$, $Mdn = 3.50$).

Examining each item individually, the results showed that the fourth heuristic and guideline were rated very highly on amount of detail, but very low in terms of ease of use and confidence. This suggested that this fourth heuristic and guideline need improvement. The heuristic and guideline concerned the phrasing of the policy, creation suggestions, and error message statements; and since the same principles were applied to all three features, the decision was made to combine them in one heuristic and guideline. However, this seemed overwhelming for the experts, as one of them commented: '*Could this be split into different heuristics and guideline? It seems like*

there are too many different things in a single heuristic and guideline (e.g. it is about policies, suggestions and error messages all at once). It seems a bit hard to measure. In terms of use it may be a bit long and complicated for one heuristic and guideline' (Expert 4). Therefore, in the fourth (and final) version of the *PassHeuristics* and *PassGuidelines*, the fourth heuristic and guideline in the third version were divided into two separate heuristics and guidelines: one addressing the policy statement, and the other addressing the suggestions and error messages. Since both of them covered the statement construction, there was some overlap between them. Finally, the fourth (and final) version of the *PassHeuristics* and *PassGuidelines* contained 10 items.

10.2.4 Final Version of the *PassHeuristics* and *PassGuidelines*

The final *PassHeuristics* and *PassGuidelines* are presented below in Table 10.19 and Table 10.20, respectively.

Table 10.19 *PassHeuristics* to support the evaluation of PCSs

Code	Heuristic
<i>Heuristic 1</i>	Provide at least two of the following features to support the user in the password creation process: password policy, password creation suggestions, and an indication of password strength (e.g. a password meter).
<i>Heuristic 2</i>	Provide the password policy before the users start entering their password and keep it available to them during and after password entry.
<i>Heuristic 3</i>	Provide password creation suggestions only once password entry has begun and leave it available during password entry and after entry is completed.
<i>Heuristic 4</i>	Provide a clear, sufficiently detailed and logical statement of password policy.
<i>Heuristic 5</i>	Provide a clear, sufficiently detailed and logical statement of password creation suggestions and error messages.
<i>Heuristic 6</i>	Provide basic information about required password composition in the policy
<i>Heuristic 7</i>	Make careful use of dynamic presentation of the password policy and password creation suggestions.
<i>Heuristic 8</i>	If possible, provide information about the strength of the password.
<i>Heuristic 9</i>	Provide clear indication of whether a password is valid or not.
<i>Heuristic 10</i>	Include password re-entry and positive confirmation.

Table 10.20 *PassGuidelines* to support the development of PCSs

Code	Guideline
<i>Guideline 1</i>	The system should provide at least two of the following features: password policy, password creation suggestions, and an indication of password strength (e.g. a password meter).
<i>Guideline 2</i>	The system should always provide the user with a statement of the password policy as soon as they open the system, and the information should remain visible throughout the password creation process.
<i>Guideline 3</i>	The system may provide password creation suggestions as soon as the user starts typing in their password, and the information should remain visible throughout the password creation process.
<i>Guideline 4</i>	<p>The password policy should be stated in specific, clear and easy to understand terms. Provide enough detail without overwhelming the user with too much information.</p> <p>Make sure that the logic of statements, such as what constitutes a valid and invalid password, is completely clear.</p> <p>Be consistent in terminology and avoid ambiguous or jargon terms.</p> <p>The password policy should be given in a declarative form before password entry (e.g. The password needs to include at least six characters), but in procedural form during and after password entry (e.g. Include at least six characters).</p>
<i>Guideline 5</i>	<p>The creation suggestions and error messages should be stated in specific, clear and easy to understand terms. Provide enough detail without overwhelming the user with too much information.</p> <p>Be consistent in terminology and avoid ambiguous or jargon terms.</p> <p>Password creation suggestions and error messages should be in declarative form throughout the password creation process (e.g. You can improve your password by having a combination of numbers, letters and symbols like !@#~)</p>
<i>Guideline 6</i>	The system should provide information regarding the minimum length of valid passwords (and the maximum length, if applicable) and the required character classes (e.g. numbers, capital letters, symbols).
<i>Guideline 7</i>	<p>When using a dynamic presentation of the password policy, make sure this will be easily understood by users and not distract them. An example of dynamic presentation is when elements required by the policy are displayed before the user starts entering his password, and that display then changes as elements are included in the password.</p> <p>Do not remove an element required by the policy from the display when the user's password includes that element.</p>

	Do not only use colour to indicate that an element has been included. Good practice is to change colour (e.g. from red to green) and indicate that the element has been included with a tick.
<i>Guideline 8</i>	It is helpful if the system can provide information about the strength of the password (e.g. a password meter), why the password has been assigned this level of strength, and how to make it stronger. Strength should be indicated both graphically (e.g. with a three-colour, traffic light metaphor) and in text.
<i>Guideline 9</i>	Clearly differentiate between an invalid password and a weak but valid password.
<i>Guideline 10</i>	Ask the user to re-enter her password for validation. Provide positive confirmation that the two entries match, as well as feedback on the lack of a match.

10.3 Evaluation of the *PassHeuristics*

The second part of this study consisted of an evaluation of the final version of the *PassHeuristics* by usability experts. Nine usability professionals were asked to conduct usability reviews of four PCSs using the *PassHeuristics* in order to evaluate them. After they completed the evaluations, the experts were asked to rate the heuristics and provide feedback about them. However, due to time and resource constraints, the present author only evaluated the heuristics and further work should evaluate the guidelines.

10.3.1 Method

10.3.1.1 Design

A within-participants design was used in this study. Each expert evaluated four PCSs using the *PassHeuristics*. Experts were allowed up to 15 minutes to conduct the evaluation of each PCS. They were given the option to move to the next PCS if they finished before the allocated time was up. They were asked to study the *PassHeuristics* and then the heuristics were available for them to consult during the evaluation for each PCS.

During the evaluation, experts were asked to identify usability problems, specify the heuristics being violated, and indicate the severity ratings on a 5-point Likert item from 1 = ‘very minor’ to 5 = ‘very major’. When specifying violated heuristics, experts had the option of not choosing any heuristic if they thought the problem was not covered by the *PassHeuristics*. In addition, they also had the option of specifying more than one heuristic if they thought the problem was related to more than a single heuristic.

After conducting the four evaluations, a questionnaire was used to measure the following variables regarding the *PassHeuristics*: (1) perceived ease of understanding, (2) perceived clarity, (3) perceived amount of detail, (4) perceived ease of use, (5) perceived usefulness, and (6) the respondents’ confidence in evaluating a PCS using these heuristics. The survey also collected information about the experts’ overall opinion of the *PassHeuristics* by measuring (7) the intention to use and (8) the perceived completeness. Each measure in the survey was formulated using a 5-point Likert item (higher = better).

10.3.1.1 Experts

Nine usability experts participated in the study, the majority of whom worked or studied at the University of York. Three were women and five were men, and their ages ranged from 23 to 66 with a mean age of 39.50 years (standard deviation = 33.00). On average, experts had 11.25 years’ experience in usability. Half of the experts (5, 55.56%) were native speakers of English, whereas the remaining had been speaking English for 20.25 years (standard deviation = 5.80). All had a postgraduate degree. The experts were entered in a prize draw of 10 Amazon vouchers worth GBP 20 each.

10.3.1.2 PCSs

The four PCSs used in this study were exactly the same as those used in Study 4 (see Table 6.2, Section 6.2.3.1.1 in Chapter 6). These PCSs contained more than 50% of the usability problems users encountered in Study 3 (see Chapter 5). The four PCSs

were *Mockup1-Apple*, *Mockup2-DailyMail*, *Mockup3-Netflix*, and *Mockup4-WordPress*.

10.3.1.3 Materials

A web-based application and a questionnaire were developed to conduct the study. The application was used for the evaluation task (henceforth called evaluation application), whereas the questionnaire was used for the questionnaire. Figure 10.3 illustrates the overall structure of the evaluation application.

The application started with the homepage, which explained the overall purpose of the study and provided an informed consent form. Next, the *PassHeuristics* were presented. Afterwards, an instruction page was provided to explain the evaluation task that the experts were about to start.

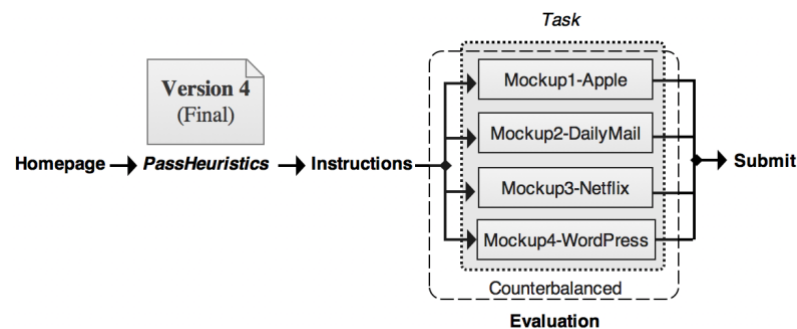


Figure 10.3 Structure of the evaluation application

Four evaluation pages were developed, one for each of the PCSs. Figure 10.4 shows an example of one of the four pages, the *Mockup3-Netflix* page. The order of presenting these pages was counterbalanced between experts. The page was divided into three sections: the PCSs, a form to indicate usability problems, and the full *PassHeuristics*. The PCS section was at the top of the page, and presented the mock-up PCSs. Experts had the chance to try successful and unsuccessful passwords and see how the PCS responded to those passwords. Next, the usability problem form was in the middle of the page. The form consisted of three elements: a text box describing the usability problems, drop-down lists for the severity ratings, and the heuristic numbers.

Every time the experts added a usability problem, the entry was shown in a list of problems table. After that, the full list of heuristics was presented for experts' reference.

Password Creation Systems Heuristics (PassHeuristics)

Time Available

Password Creation System

Create New Password

Account Name

Choose a password (12-20 characters, at least one uppercase letter, at least one lowercase letter, at least one number or symbol)

Adding usability problem

Description of the problem	Severity of the problem	Violation of heuristic
<input type="text"/>	<input type="text"/>	<input type="text"/>
<input type="button" value="Add"/>		

List of added usability problem

No.	Description of the problem	Severity of the problem	Violation of heuristic

Heuristic 1. Provide at least two of the following features to support the user in password creation: password policy, password creation suggestions, an indication of password strength (e.g. a password meter).

Heuristic 2. Provide the password policy before the users start entering their password and leave it available to them during password entry and after entry is completed.

- The system should always provide the user with a statement of the password policy as soon as they open the system, and the information should remain visible throughout the password creation process.

Heuristic 3. Provide password creation suggestions only once password entry has begun and leave it available during password entry and after entry is completed.

- The system may provide password creation suggestions as soon as the user starts typing in their password, and the information should remain visible throughout the password creation process.

Heuristic 4. Provide a clear, sufficiently detailed and logical statement of password policy.

- The password policy should be stated in specific, clear and easy to understand terms. Provide enough detail without overwhelming the user with too much information.
- Make sure that the logic of statements, such as what constitutes a valid and invalid password, is completely clear.
- Be consistent in terminology and avoid ambiguous or jargon terms.
- The password policy should be given in a declarative form before password entry (e.g. The password needs to include at least six characters), but in procedural form during and after password entry (e.g. Include at least six characters).

Heuristic 5. Provide a clear, sufficiently detailed and logical statement of password creation suggestions and error messages.

- The creation suggestions and error messages should be stated in specific, clear and easy to understand terms. Provide enough detail without overwhelming the user with too much information.
- Be consistent in terminology and avoid ambiguous or jargon terms.
- Password creation suggestions and error messages should be in declarative form throughout the password creation process (e.g. You can improve your password by having a combination of numbers, letters and symbols like !@# -)

Figure 10.4 A screenshot of the *Mockup3-Netflix* evaluation page

Upon completion of the evaluation task, experts were directed to the post-task questionnaire. The questionnaire consisted of three parts:

- The first part asked experts for their opinion on individual heuristics by asking them to rate the (1) perceived ease of understanding, (2) perceived clarity, (3) perceived amount of detail, (4) perceived ease of use, (5) perceived usefulness, and (6) their confidence in evaluating a PCS using this heuristic. These variables were measured using a 5-point Likert item ranging from 1 (not at all easy to understand/ not at all clear/ far too little detail/ not at all easy to use/ not at all useful/ not at all confident) to 5 (extremely easy to understand / extremely clear/ far too much detail/ extremely easy to use/ extremely useful/ extremely confident). An optional open-ended question also gave participants the chance to explain their ratings.
- The second part asked the experts about their opinion on the overall *PassHeuristics* by asking them to rate their intention to use them and the perceived completeness. These variables were measured using a 5-point Likert item ranging from 1 (not at all likely/ not at all complete) to 5 (extremely likely/ extremely complete).
- The third part asked questions about participants' demographic characteristics: age, gender, native language, education, and experience in the usability field.

10.3.1.4 Pilot of the Study Procedure

A pilot study was conducted with one usability expert. The expert found the overall procedure easy to follow and the instructions clear. However, the expert did raise a concern regarding the evaluation task: she noted that there were two questions about the violation of the heuristic for each usability problem. These were (1) if the expert thought that more than one heuristic was violated, and (2) if the usability problem was not covered by the proposed heuristics. Therefore, in the main study, experts were given the option to choose more than one heuristic to address (1); and they also were

given the choice of not selecting any heuristic in case of (2). The data from the pilot session were not included in the data analysis.

10.3.1.5 Procedure

Links to the study were posted/emailed via social networks to the following specialised usability groups: User Experience Professionals Association (UXPA) Group, Special Interest Group on Computer–Human Interaction (SIGCHI Group), City University Centre for HCI Design Group, British Computer Society (BCS) Interaction Specialist Group, User Experience, User Experience and Human-Computer Interaction (UX/HCI) researchers, BCS Interaction Group, and Human-Computer Interaction (HCI) Group at the University of York.

A briefing about the study and an informed consent form were provided at the beginning of the application. Experts were assured that they would not be asked to reveal any of their passwords, and that even the passwords that they tried during the study would not be stored. Experts then confirmed their agreement and their understanding of the information provided in the briefing by clicking on the ‘Next’ button.

After that, experts were asked to read and carefully study the full *PassHeuristics*. Once they were finished, they were given the task instructions for the evaluation. Each expert then evaluated four mock-up PCSs, and the order of presentation was counterbalanced.

For each mock-up PCS, the expert completed only one task: creating a new password with as many possibilities as he wished to see the how the PCS behaved. For each potential usability problem, the expert was asked to describe it, rate its severity using a 5-point Likert item, and finally select which heuristic was being violated. If the expert thought the usability problem was not covered by the list of heuristics, she chose ‘none’. This procedure of identifying usability problem was repeated until the expert

felt there were no more problems to be identified in the mock-up PCS, at which point he clicked on the ‘Done’ button and moved to the next mock-up PCS.

Upon completion of the evaluation, experts were directed to the post-tasks questionnaire to answer questions about their opinion on the individual heuristics, the overall *PassHeuristics*, and finally, some demographic information.

10.3.1.6 Data Analysis

Since the second aim of this study was to investigate the effectiveness of the *PassHeuristics*, not all collected data was analysed. Instead, the analysis included only the experts’ opinions about the *PassHeuristics* in the questionnaire after they had completed the evaluation.

Kolmogorov-Smirnov and Shapiro-Wilk tests were used to test for normality on all dependent measures used in the post-tasks questionnaire. The majority of dependent measures were significantly non-normal ($p < 0.05$) for both tests. Therefore, nonparametric statistics were used throughout the analysis. Friedman’s test was used for each measure to compare the ratings between the 10 items of the *PassHeuristics*.

10.3.2 Results

In total, 110 usability problems were identified using the *PassHeuristics*, with a mean of 12.22 per PCS (standard deviation = 8.60). Table 10.20 summarises the results for the dependent measures for each heuristic after the experts used them to evaluate the four PCSs.

The results showed no significant difference between the individual heuristics in the ratings of ease of understanding ($\chi^2(9) = 9.45$, $p = .398$), clarity ($\chi^2(9) = 11.83$, $p = .223$), amount of detail ($\chi^2(9) = 13.99$, $p = .122$), ease of use ($\chi^2(9) = 13.79$, $p = .130$), usefulness ($\chi^2(9) = 9.63$, $p = .381$), or confidence ($\chi^2(9) = 9.51$, $p = .392$).

Table 10.21 Mean (median) ratings of the six dependent measures for individual items of the *PassHeuristics*

	Ease of understanding	Clarity	Amount of detail	Ease of use	Usefulness	Confidence
#1	3.33 (4.00)	3.78 (4.00)	2.78 (3.00)	3.00 (3.00)	3.33 (3.00)	3.67 (4.00)
#2	4.00 (4.00)	4.00 (4.00)	3.00 (3.00)	3.67 (4.00)	3.56 (3.00)	3.67 (4.00)
#3	3.75 (4.00)	3.63 (4.00)	2.88 (3.00)	3.50 (3.50)	3.38 (3.00)	3.50 (3.00)
#4	3.38 (3.00)	3.50 (3.00)	3.38 (3.00)	3.38 (3.00)	3.63 (3.50)	3.00 (3.00)
#5	3.75 (3.50)	3.63 (3.50)	3.13 (3.00)	3.38 (3.00)	3.63 (3.50)	3.63 (3.50)
#6	4.13 (4.00)	4.25 (4.50)	2.88 (3.00)	4.25 (4.50)	4.00 (4.00)	3.88 (4.00)
#7	3.50 (3.50)	3.50 (3.50)	3.00 (3.00)	3.25 (3.00)	3.25 (3.00)	3.50 (3.00)
#8	4.25 (4.00)	4.25 (4.00)	3.00 (3.00)	4.00 (4.00)	4.00 (4.00)	3.63 (3.50)
#9	4.00 (4.00)	3.88 (4.00)	2.63 (3.00)	3.75 (4.00)	3.88 (4.00)	3.75 (4.00)
#10	3.88 (4.00)	4.00 (4.00)	2.75 (3.00)	3.88 (4.00)	3.88 (4.00)	3.63 (3.50)
<i>p value</i>	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
<i>Overall</i>	3.68 (3.60)	3.81 (3.50)	2.90 (3.00)	3.50 (3.30)	3.63 (3.40)	3.53 (3.40)

A one-sample Wilcoxon Signed Rank test was used for each measure to compare the observed median against the hypothesised median (midpoint = 3) across the individual heuristics and the overall *PassHeuristics*. The results indicated the following:

- Ease of understanding: the ratings of ease of understanding were significantly higher than the midpoint for the second ($Z = 41.00$, $p = .021$), sixth ($Z = 21.00$, $p = .024$), eighth ($Z = 36.00$, $p = .008$), ninth ($Z = 32.50$, $p = .033$) and overall heuristics ($Z = 41.00$, $p = .028$).
- Clarity: the ratings of clarity were significantly higher than the midpoint for the second ($Z = 41.00$, $p = .021$), sixth ($Z = 21.00$, $p = .023$), eighth ($Z = 36.00$, $p = .008$), tenth ($Z = 32.50$, $p = .033$), and overall heuristics ($Z = 45.00$, $p = .007$).
- Ease of use: the ratings of ease of use were significantly higher than the midpoint for the sixth ($Z = 21.00$, $p = .023$) and eighth heuristics ($Z = 28.00$, $p = .011$).

- Usefulness: the ratings of usefulness were significantly higher than the midpoint for the sixth ($Z = 15.00$, $p = .038$), eighth ($Z = 28.00$, $p = .011$), and overall heuristics ($Z = 36.00$, $p = .012$).
- Confidence: the ratings of confidence in using the heuristic were significantly higher than the midpoint for the sixth ($Z = 15.00$, $p = .038$) and overall heuristics ($Z = 15.00$, $p = .043$).

All in all, the experts rated their likelihood of using the *PassHeuristics* in the future as very likely ($M = 3.75$, $Mdn = 4.00$). Furthermore, they perceived the *PassHeuristics* to be complete ($M = 4.00$, $Mdn = 4.00$).

10.4 Discussion

Following the first two phases of this research, the present study aimed to (1) develop a set of usability heuristics and guidelines for use in guiding the evaluation and development of PCSs, and (2) to evaluate the proposed heuristics. To address these aims, this study had two parts: (1) the development of the proposed heuristics and guidelines (2) their evaluation.

To develop the proposed heuristics and guidelines, a bottom-up approach was employed in this study since it was grounded by real-world data reflecting real users' usability problems. Hence, the proposed heuristics were created on the basis of usability problems encountered by users and supported by user perception and performance data collected in this research. The proposed heuristics and guidelines consist of 10 items that cover more than 75% of the usability problems identified by users.

For better use in evaluating systems, the heuristics should be short and understandable. As for the number of items included in the heuristics, most of the well-used sets of heuristics, such as those by Nielsen (1994), Shneiderman (2018), and Norman (2002), contain between 7 and 10 items. Thus, the *PassHeuristics* fit well with these heuristics. Furthermore, three rounds of feedback were performed with a total of six usability

professionals to make sure the heuristics and guidelines were understandable even if they were designed specifically for PCSs. Overall, the proposed set of heuristics and guidelines were perceived to be easy to understand and clear.

Subsequently, nine usability professionals evaluated four mock-up PCSs to evaluate the *PassHeuristics*. The results of the evaluation indicated that experts made good use and interacted well with the heuristics, as they identified in total 110 usability problems, with a mean of 12.22 per PCS.

Overall, the results revealed that the *PassHeuristics* were perceived to be easy to understand, clear, and useful. Evaluators also felt confident evaluating a PCS using this set of heuristics. In addition, they expressed a high level of intention to use the *PassHeuristics* in the future, as they covered all aspects of PCSs. Although the ratings on the dependent measures of the individual heuristics did not differ between the different items, it was interesting to examine each heuristic individually across the six measures (see Table 10.20).

The first heuristic concerns the provision of the supporting features. Contrary to expectations, it did not receive significantly high ratings across the six dependent measures, although it was on average above the midpoint. It is difficult to explain this result as the current data do not give further explanation for this result.

The second heuristic, regarding the timing of presentation of the statement of policy, was rated significantly highly in the level of ease of understanding and clarity. Therefore, it is somewhat unexpected that the third heuristic, which addresses the same point but for the statements of suggestion, was not rated as highly as the second one.

The fourth and fifth heuristics concern the phrasing of statements of password policy and creation suggestions, respectively. Interestingly, one evaluator commented that using these heuristics was difficult since the consequence of violating them was not clearly seen: *‘it is difficult to judge to what extent a password system is violating it.*

Whether it is just a cosmetic issue of not changing wording (declarative vs procedural) or whether it is too long and overwhelming’ (Expert 3).

The evaluators rated the sixth heuristic, regarding the password composition and requirements, significantly highly on almost all six dependent measures (except amount of detail).

The seventh heuristic, which concerns the use of dynamic presentation, did not receive significantly high ratings for any of the six dependent measures. This might be due to the evaluators’ contradictory views about this particular heuristic when they used it in the evaluation. For instance, one evaluator noted how well this heuristic was explained, while another thought one of its aspects made no sense.

The eighth heuristic, regarding the design of the strength indicator, was rated significantly highly in terms of ease of understanding, clarity, ease of use, and usefulness.

The evaluators rated the ninth heuristic, which covers the necessity of providing feedback about the validity of the password, significantly highly in the level of ease of understanding only.

Finally, the tenth heuristic concerns the provision of password re-entry field associated with positive confirmation. It was rated significantly highly in terms of clarity. Interestingly, one evaluator commented that the positive confirmation given to users should be optional: *‘It is mandatory to post positive feedback when the passwords match. I feel that the common practice is you assume users enter matching passwords and feedback is given only when they don’t match’ (Expert 1).*

It is interesting to note that none of the usability professionals who took part in this study (except the internal expert) engaged in any research within the scope of PCSs, and they had only just encountered these heuristics. However, they intended to use them in the future, as they felt they covered all aspects of PCSs and were perceived as

easy to use and useful. This suggests that the *PassHeuristics* are a useful tool to support the evaluation of PCSs.

These findings must be interpreted with caution as the generalisability of these results is subject to certain limitations. First, an online recruitment method was used targeting eight specialised usability groups, but relatively a small number of evaluators responded and took part. Second, this study evaluated the proposed heuristics using self-report data. Because of the aims of the study (development and evaluation), we did not examine the usability problems reported by the experts. Analysing these would give us a greater understanding of the effectiveness of the heuristics. In particular, further analysis is needed to investigate whether the 110 usability problems identified by the experts in this study are the problems users had encountered.

10.5 Conclusions

It is important to use appropriate usability heuristics and guidelines when conducting usability evaluations or designing interactive systems. Existing sets of usability heuristics do not cover the problems associated with password creation in any detail. All in all, the proposed *PassHeuristics* and *PassGuidelines* developed specifically for the evaluation and development of PCSs should help evaluators and developers in their work. Evaluators in this final study intended to use *PassHeuristics* in the future, as it covered all aspects of PCSs and was perceived as easy to understand, clear and useful. The *PassHeuristics* and *PassGuidelines* can contribute to the improvement of the password creation process and consequently to the security of digital data and peace of mind for owners of such data.

Chapter 11

General Discussion and Conclusions

This research aimed to inform the design and usability of PCSs and their supporting features so they can better support users when creating passwords. This thesis addressed this specific goal at the user interface level by providing knowledge about how users react to a range of aspects of the supporting features in PCSs and by providing a set of usability heuristics and guidelines that support the evaluation and development of existing PCSs.

The central research question in this thesis was, *how can PCSs effectively support users in creating passwords without compromising security?* The thesis answered this question by breaking the research into three phases: (1) understanding the current practices of PCSs, (2) testing design variables and examining their effects on password creation and users, and finally (3) proposing usability heuristics and guidelines specifically for the evaluation and design of PCSs. Each phase consisted of studies that provided insights for the next one. The first phase consisted of Studies 1, 2, 3, and 4 (see Chapter 3, 4, 5, and 6), the second phase consisted of Studies 5, 6, and 7 (see Chapter 7, 8 and, 9), and the third phase consisted of Study 8 (see Chapter 10). A mixture of qualitative and quantitative methods was adopted to address the research aim.

The following section summaries the main findings of each phase and subsequent section discusses the implications of the findings.

11.1 Overall Summary

11.1.1 Phase 1: Understanding the Current Practices of PCSs

The aim of this phase was to provide a better understanding of the current practices of PCSs. A total of 30 current PCSs was examined (see Study 1, Chapter 3). The analysis of PCSs showed that there is little consistency among PCSs. The findings of this analysis also revealed that the support provided by current PCSs does not seem adequate for users choosing a password. This may create usability issues for users, as they cannot predict how a new PCS will work when they encounter it.

To investigate the usability of PCSs further, both expert (see Study 2, Chapter 4) and user (see Study 3, Chapter 5) usability evaluations were conducted on a number of PCSs from the set already analysed in the first study. The two evaluations produced a pool of 121 distinct usability problems: 40 (33.06%) found by experts only, 38 (31.40%) by users only, and 43 (35.54%) by both experts and users. The nature of the usability problems fell into one of six main categories: *password policies*, *password creation suggestions*, *password strength indicators*, *other feedback*, *error messages*, and *other* problems. Focusing on the three main supporting features, the usability evaluations generally indicated that problems were related to the lack of supporting features and the design presentation of these features. Moreover, problems were related to the amount of detail and clarity of instructions for creating passwords.

To have a comprehensive understanding of current PCS practices, it was important to examine the impact of these practices and usability problems on users and the passwords they generate (see Study 4, Chapter 6). The main findings revealed that current PCS practices had different effects on PCS usability and password strength. However, it was difficult to determine a specific practice that might have caused this effect, as there was a high level of interaction between the components integrated at the user interface level.

All in all, this phase contributed to the overall research by providing an understanding of the problems that people encounter when creating passwords. A corpus of usability

problems with PCSs was collected through user and expert evaluations. The evidence from this phase clearly suggests the main interface flaws in current PCSs to be further investigated and fixed.

11.1.2 Phase 2: Testing Design Variables for PCSs

The aim of this phase was to investigate different aspects of the design of PCSs. The design aspects and variables to examine in this phase were based on the outcomes of the studies in the previous phase.

For the user instructions for creating passwords (see Study 5, Chapter 7), the users' perceptions of the most frequently used instructions in current PCSs were examined. One of the main findings showed that the commonly used instructions of password policy and creation suggestion in PCSs vary widely and do not match users' needs for creating passwords.

In terms of the design aspects of each supporting feature (see Study 6, Chapter 8), the following variables were examined: (1) timing of presentation for both the policy and creation suggestions, and (2) media and colour-scheme for the strength indicators. In general, the findings suggested that different design presentations of the supporting features affected the PCS usability and password strength differently when users created passwords, but not when they recalled them. Another finding showed that in general the mere presence of supporting features affected the usability of the PCS and password strength during both the creation and recall processes.

In terms of the interaction between the supporting features (see Study 7, Chapter 9), the effects of presenting more than one supporting feature in a PCS were examined. The study used the outcomes identified in Study 6 for each supporting feature to design four combinations of supporting features. The findings revealed that different combinations affected the PCS usability and the password strength differently when users created passwords, but not when they recalled them. In addition, the mere presence of combined supporting features affected the PCS usability and password strength only when users created passwords.

Given all these findings, this phase contributed to the overall research by providing an understanding of user instructions in the field of password research. It also improved the understanding of how PCSs should be designed their supporting features to improve usability and password strength. Finally, these studies confirmed previous findings and provided additional evidence that the presence of supporting features affects PCS usability and password strength.

11.1.3 Phase 3: Proposing Usability Heuristics and Guidelines for PCSs

Based on the findings of the previous two phases, a set of usability heuristics and guidelines was proposed for the evaluation and development of PCSs (see Study 8, Chapter 10). The proposed guidelines and heuristics, *PassHeuristics*, contain 10 heuristics that cover more than 75% of the usability problems identified by users. These guidelines and heuristics are grounded in empirical data, specifically from the usability problems users experienced and their perceptions of current PCSs (Study 3), in addition to supporting evidence from the experimental data (Studies 5, 6, and 7). The evaluation revealed that the *PassHeuristics* were perceived to be easy to understand, clear, and useful. Furthermore, the evaluators intended to use the *PassHeuristics* in the future, as they cover all aspects of PCSs.

11.2 Discussion

Users' behaviours that are linked to the password problem can be seen from two perspectives: the system side (see Section 11.2.1) and the user side (see Section 11.2.2).

The system side examines when users create their passwords and how they do so by studying their password choice and the PCSs they use at the user interface level. This thesis focused mainly on the system side as it was the main interest of the present author to study use of PCSs at the user interface level. The central research question of this research is answered in the system side section.

On the other hand, the user side examines how users handle their passwords by understanding their password-related behaviours. This research collected self-reported data about users' password-related behaviours in password creation and management. A total of 829 participants took part in the online user studies (Studies 4, 5, 6, and 7), and answered questions about their password-related behaviours. The reported data across the four studies was generally consistent; thus, the overall implications of these studies are discussed in the user side section.

11.2.1 Implications from the System Side Perspective

Designing secure and usable PCSs is only one step towards protecting the users' assets online. One of the key design principles that should be adopted for designing the user interfaces is consistency, thus, it is very important to ensure that all PCSs on different websites are as consistent as possible. Consistent PCSs will not only help users to create usable and secure passwords, but also will help designers and developers of PCSs to avoid poor design decisions. This thesis has provided guidelines for developers and heuristics for evaluators of PCSs as tools to help in the creation of secure but usable PCSs. Although the proposed heuristics and guidelines were constructed from the users' perspective, both usability and security aspects were considered throughout their development. Thus, security aspects are already embedded in the heuristics and guidelines which ensure that security will be considered by the evaluators and developers. Thus, the proposed heuristics and guidelines to some extent balance the trade-off between the usability and security. Furthermore, providing developers with a set of guidelines would help them to design consistent PCSs, that eventually support the current trend in the industry for maintaining consistency.

This research showed that most current PCS practices are implemented in inconsistent manners; there was no standard practice in employing the supporting features in such a way that provided consistency which would have helped and supported users. One of the problems caused by this inconsistency is that PCSs have different designs of these features that do not seem adequate, which consequently lead to probable user confusion and ambiguity. For example, some PCSs implement a password strength

indicator in their user interface but do not tell users how to increase their password strength or why the chosen password is weak. Another example is that some PCSs offer both a statement of password policy and a creation suggestion without being careful about the language used to clearly distinguish them; this may easily cause user confusion regarding what is mandatory and what is optional. When participants in the online studies was asked about their experience with current PCSs, around 14.19% users reported having a negative previous experience with a PCS which made them leave the website and disrupt their primary task. One of their frustrations was related to difficulty in complying with the given policy, which took them a long time.

Although PCSs are small interactive systems, having a conceptualised model of the password creation process was of great benefit to this research. The proposed three-step model of PCSs helped in understating the user interface of these systems in a coherent and clear way. The model included information about the interaction steps and supporting features which may be available at each step in a PCS. Identifying these steps and supporting features helped in evaluating the PCSs with experts and users, along with categorizing the usability problems. Furthermore, it provided a starting point to determine the design aspects to be examined in PCSs. It may be the case therefore that the three-step model provides a foundation for both the user interface design and the evaluation of PCSs.

The research findings of this thesis suggest that PCSs can effectively support users in creating passwords by addressing four key factors: (1) provision of supporting features, (2) user instructions for creating passwords, (3) timing of presentation for presenting statements of policy and creation suggestions, and finally (4) media and colour scheme for designing strength indicators. The following discusses each factor from a system perceptive.

(1) Provision of supporting features

The use of password policy as a supporting feature was by far the most common in PCSs. Current PCSs offer policy either as an individual feature or combined with other features. The findings in this research showed that when the policy, creation

suggestions, and strength indicator (for both multi-colour and single-colour) were provided as individual features or combined with others, users were less efficient in creating passwords yet more satisfied than when no features were provided at all. However, users created stronger passwords when the PCSs offered the policy alone, the strength indicator alone (with multi-colour), or combined features. This outcome clearly suggests the importance of providing supporting features, as they improve password strength and user satisfaction at the expense of efficiency in completing the task.

It was also found in the analysis that current PCSs always offer the strength indicator feature combined with another one, but not on its own. Interestingly, the findings of this research showed that the mere presence of the strength indicator improved password strength when compared to not providing any features at all. However, this effect was only found with multi-colour indicators. It seems that the use of strength indicators could be a potential replacement for the traditional way of making users create strong passwords (i.e. the use of policy) as long as careful consideration is taken in designing them.

These findings of this study confirm those of previous studies that have examined the effect of providing a policy statement (Campbell et al., 2011; Proctor et al., 2002; Vu et al., 2007) and a strength indicator on password strength (e.g. Furnell & Esmael, 2017; Ur et al., 2012).

Regarding the interaction between these features, the findings revealed that different combinations had an effect on the efficiency of PCSs and the strength of passwords. However, each combination affected password characteristics differently. For example, users chose passwords with a high number of digits but low number of lowercase letters when the PCS presented them with the policy statement in combination with the creation suggestion. Furthermore, the different combinations did not affect the level of user satisfaction: users were not more satisfied with one combination than another.

One might expect that providing an extra supporting feature during the password creation process would improve user satisfaction regardless of the type of features combined, but this was not true. This led to the question of whether one supporting feature was sufficient without overloading users with a great deal of information during the password creation process. In this vein, a comparison between presenting an individual feature and combined features yielded interesting findings: users were more satisfied when a supporting feature was combined with another one than when it was presented alone, without affecting their efficiency.

(2) User instructions for creating passwords

More than half of the users in the online studies (55.37%) reported not always reading user instructions when creating passwords. Common reasons were given for not doing so; these reasons related to the legibility, length, and visibility of provided instructions. The findings from the usability studies supported what the users reported: 60.00% of the usability problems identified regarding password policy were related to the amount of detail and clarity of the instructions, and the same was true of 35.00% of problems regarding creation suggestions.

The analysis of the commonly used instructions of password policy and creation suggestion in PCSs vary widely and do not match users' needs for creating passwords. Users preferred declarative policy before they interacted with a PCS, but procedural policy during and after the interaction. For creation suggestions, users preferred declarative statements before, during, and after interaction with a PCS. Therefore, the combined qualitative and quantitative evidence suggests why users have difficulty choosing secure passwords, since user instructions play a key role in understanding password requirements.

(3) Timing of presentation for presenting statements of policy and creation suggestions

Since creating a password is a secondary task for users and efficiency is a crucial aspect to consider, users encountered usability problems related to the presentation of the policy and suggestion later than expected. Investigating this further, the findings

showed that the different timings of presentation had an effect on the efficiency of PCSs. The password compliance rate was also affected by each timing of presentation.

The analysis of PCSs showed that current PCSs offer statements of password policy with almost the same frequency across the three main timings of presentation (i.e. before, during, and after password entry), while they mainly offer suggestions before password entry. It seems that the timing these features' presentation is mainly ignored in current PCSs. Therefore, PCS designers should pay more attention to this aspect, as it not only affects the efficiency with which users can create passwords with the system but also the strength of the passwords created with the system. Users created policy-compliant passwords efficiently and had a high successful recall rate when the statement of policy was presented before password entry. Furthermore, they created longer passwords containing all character classes more quickly, and were more confident when the statement of suggestion was presented during password entry than the other timings of presentation.

(4) Media and colour scheme for designing strength indicators

There are a number of different possible designs for password strength indicators. Two design attributes were identified for investigation in this research: the media used and the colour scheme. Both multi-colour and single-colour coding schemes were encountered in the strength indicators, with the majority of the PCSs using the former. Some of the PCSs used the "traffic light" metaphor to indicate the strength of passwords. On the other hand, very few PCSs provided a single-colour indicator. The usability evaluations revealed usability problems that related to the colour contrast and coding schemes used in current PCSs.

The two attributes showed interesting implications for designing strength indicators when they were investigated. For example, providing a graphical indicator without explaining what the changes in the bar mean may result in weaker passwords and poor usability. In addition, using only one colour to distinguish between the strength levels may also result in weaker passwords and poor usability, whereas using the traffic light metaphor of green, amber, and red colours results in stronger passwords. Users created

stronger passwords that were longer and contained four character classes, and they were more satisfied when the strength indicator was presented with both graphical and textual presentation using the traffic light metaphor than in other conditions.

11.2.2 Implications from the User Side Perspective

In this research, users reported that they had on average around 21.71 password-protected accounts (standard deviation = 9.99) and approximately 11.99 passwords (standard deviation = 3.58). Interestingly, this finding is to some extent in agreement with that of Florencio and Herley (2007), who found that each person had about 25 accounts and 7 passwords. That data was collected in 2006, so one would expect this figure to be considerably higher now; however, this is not the case. It may be that users underestimate the number of password-protected accounts they have. On the other hand, the number of passwords has increased even though the number of password accounts did not. A possible explanation for this might be that users have become more cautious and are not reusing the same passwords. However, this explanation is not valid either. Users in this research reported that they still used the same (71.17%) or slightly different passwords (67.79%) for multiple accounts.

The users had a good understanding of what makes a secure password. They commented that secure passwords should have a combination of different character classes, and that they should not be based on personal information that might make them easy to guess. Furthermore, others mentioned the length and the use of password managers as criteria for making secure passwords. All in all, 38.48% of users described themselves as being very knowledgeable about creating secure passwords, and 36.67% felt very confident about the strength of their most complicated password.

Although there is little evidence in the literature about users being unlikely to change a password once it is set, in this research 39.45% users reported changing their passwords every three to six months, while 6.88% never did. The former is an unexpectedly high figure, but it might be due to the increasing number of security breaches that have happened in recent years.

Regarding password management strategies, participants mentioned writing down passwords, using a password manager, reusing the same passwords for multiple accounts, modifying different variations of the same passwords, relying on their memory, and choosing passwords that were easy to remember. Furthermore, they reported different ways of keeping their passwords safe when they chose to write them down, such as using notepads, sticky notes, and encrypted files on their computer. These coping strategies match those observed in earlier studies (Brown et al., 2004; Dhamija & Perrig, 2000; Florencio & Herley, 2007; Gaw & Felten, 2006; e.g. Grawemeyer & Johnson, 2011).

11.3 Password-Related Studies

Due to the sensitivity of password data, obtaining ethical approval for password-related studies and collecting valid password data remain challenging tasks. Regarding ethical approval, the important challenge that researchers experience relates to data disclosure and confidentiality. Researchers have to make sure that participants do not reveal or use any of their own passwords in a study. Since the raw data of all studies consist of passwords, the data should be completely anonymised, and their storage strongly protected. The data files should not contain information that describes any patterns that the participants use in creating their passwords. In other words, any information that could help an attacker to easily break passwords should be avoided in password-related studies.

In terms of data collection, researchers face challenges at different levels. First, researchers have to find ways to convince participants to take part in such studies to obtain an appropriate sample size. Second, researchers have to make participants feel comfortable to share information about their password creation and management behaviours. The most challenging aspect relates to participants' perceptions about revealing information about their passwords that may identify them or their way of thinking in any way. To study recall of passwords, two-part studies are needed, with some time lapse (typically 3 – 5 days) between the two parts. This means that an

appropriate number of participants must be recruited at the first part to allow for an attrition rate of participants not returning for the second part.

One way to address challenges in data collection is the use of online studies. Unlike laboratory studies, online studies provide access to a larger pool of participants and yield high response rates. Systems such as MTurk have been regularly deployed in password-related studies (Shay et al., 2014, 2015; Ur et al., 2012). MTurk is a crowd-sourcing platform on which researchers post their tasks and participants (known as Turkers or workers) can complete these tasks in return for small payments. The use of MTurk has enabled researchers to access a wide range of the general population (Berinsky, Huber, & Lenz, 2012; Casler, Bickel, & Hackett, 2013) to conduct large studies rapidly and cost effectively. However, running password-related studies on MTurk has limitations, such as ecological validity and data quality.

Ecological validity is about whether the users' behaviour in an experiment matches a real-life situation. In general, users might be less vigilant as they are creating factitious passwords that are not for real accounts, and they are not asked about high-value accounts. The effect of this artificial situation might not only affect the password creation stage but also the recall stage. The memorability of these factitious passwords could also be affected adversely, as users would know that there are no penalties of forgetting these passwords (such locking out of an account or calling a help-desk). Thus they would tend to not worry about remembering the factitious password. However, there are ways to overcome such challenges. One way to improve the ecological validity of the task is to adopt a scenario-based approach. In this research, participants were provided with a scenario related to a bank account and were asked to imagine their needs in creating a new password in light of that scenario. Another way is the use of a recall task to make participants more vigilant about their newly created passwords, and thus improve their efforts to remember the factitious passwords.

It is crucial to maintain the quality of data over the quantity when using MTurk. Buhrmester et al. (2018) identified three factors that negatively affect the data quality on data collected using MTurk: inattention, dishonesty, and attrition. In this research, of the four studies which used participants from MTurk (Studies 4, 5, 6 and 7), the

first and last studies only excluded 8.6% and 18.5% of the participants, respectively. However, the second study excluded 48.7% of participants and the third study excluded 54.4% of the participants, which was higher than expected. The three factors suggested by Buhrmester et al. (2018) were considered in the data cleaning phase for all studies, and they occurred in the excluded cases. The attention factor was strictly checked via two measures: (1) attention-check questions (Oppenheimer, Meyvis, & Davidenko, 2009) and (2) MTurkers reputation (e.g. 95% approval rate or higher) (Peer, Vosgerau, & Acquisti, 2013). Some of the respondents gave identical answers for all questions, so it was decided to exclude them from the analysis to minimize the level of inattention in the data. For the dishonesty factor, the present author identified dishonest answers by checking the completion time. Since the average completion time was about 30 - 20 minutes for studies 5 and 6, any completion rate of less than 3 minutes was considered unreliable to include in the analysis. For the attrition factor, any incomplete responses had to be excluded to avoid unbalanced groups which affects the type of analyses to be used (Buhrmester et al., 2018). The length of the study could have resulted in a high attrition rate. For future research, it is recommended to design shorter and quicker study whenever possible like studies 4 and 7, where the exclusion rate was not high.

Self-reporting is sometimes unavoidable in password-related studies due to the sensitivity of password data. For example, information regarding whether the password is new or a modified version cannot be observed from the actual performance of the participants, but must be asked by the researcher. Although self-reporting can be inaccurate due to memorability and social desirability issues, it is the only option for the researcher in such cases of password creation and management behaviour.

11.4 Limitations and Future Work

Although this research has shown how PCSs could be improved to support users effectively when creating passwords without compromising security, it has certain limitations in terms of ecological validity. For ethical considerations of security and

privacy, and as is the case in most password studies, participants were asked to create fictitious passwords and not to use their own during the password creation process. Hence, it is difficult to assess whether the passwords obtained in the user studies reflect real passwords that users would create and use in real world situations. To improve the ecological validity, the present work used a method that is commonly employed in the literature: online studies through MTurk (Kelley et al., 2012; Shay, 2015; Shay et al., 2014; 2015). This is because MTurk has a significantly more diverse population than samples in a typical laboratory study conducted at a university (Buhrmester, Kwang, & Gosling, 2011). Other aspects were also taken into consideration to improve the ecological validity. For instance, a scenario-based approach was used during the password creation process, and the recall task was performed three days after creation.

This research served as a first step to improving the usability of PCSs as whole interactive systems, focusing on all the characteristics of the supporting features incorporated in these systems. This aim was achieved by proposing the *PassHeuristics* and *PassGuidelines* specifically for the evaluation and design of PCSs, which should hopefully contribute to the improvement of the password creation process consequently to the security of digital data and peace of mind of owners of such data. However, more information on the effectiveness of the proposed heuristics would help to establish a greater degree of validity on this matter. It would be interesting to evaluate these heuristics by using them to conduct usability evaluations on current PCSs, and comparing the results with those obtained using other usability heuristics or without any heuristics. Another possible area of future research would be applying the proposed guidelines while designing a PCS for a real website and examining its impact on the users and their created passwords.

11.5 Conclusions

User typically interact with an individual PCS only once, but interact with many PCSs on different websites multiple times. Nevertheless, the experience of a single use greatly benefits the overall usability of the website; benefits include, but are not limited to, a higher successful sign-up rate, users achieving their primary tasks quickly,

and finally, gaining user trust. Therefore, the research findings of this thesis suggest that PCSs can effectively support users in creating passwords by addressing four key factors: (1) provision of supporting features, (2) user instructions for creating passwords, (3) timing of presentation for presenting statements of policy and creation suggestions, and finally (4) media and colour scheme for designing strength indicators. The important implication of this research lays in the evaluation and design of PCSs through facilitating *PassHeuristics* and *PassGuidelines* to better support users when creating passwords.

References

- Adams, A., & Sasse, M. A. (1999). Users are not the enemy. *Communications of the ACM*, 42(12), 40–46.
- Baker, K., Greenberg, S., & Gutwin, C. (2002). Empirical development of a heuristic evaluation methodology for shared workspace groupware. *Cscw*, 96.
- Barton, B. F., & Barton, M. S. (1984). User-friendly password methods for computer-mediated information systems. *Computers and Security*, 3(3), 186–195.
- Batra, S., & Bishu, R. R. (2007). Web Usability and Evaluation: Issues and Concerns. In N. Aykin (Ed.), *Usability and Internationalization. HCI and Culture SE - 30* (Vol. 4559, pp. 243–249). Springer Berlin Heidelberg. h
- Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk. *Political Analysis*, 20(03), 351–368.
- Biddle, R., Chiasson, S., & Van Oorschot, P. C. (2012). Graphical passwords: Learning from the first twelve years. *ACM Computing Surveys (CSUR)*, 44, 19.
- Birns, J. H., Joffre, K. A., Leclerc, J. F., & Paulsen, C. A. (2002). Getting the Whole Picture: Collecting Usability Data Using Two Methods----Concurrent Think Aloud and Retrospective Probing (pp. 8–12). Presented at the Proceedings of UPA Conference.
- Bonneau, J., Herley, C., Oorschot, P. C. V., & Stajano, F. (2012). *The quest to replace passwords: a framework for comparative evaluation of Web authentication schemes*. University of Cambridge, Computer Laboratory.
- Brainard, J., Juels, A., Rivest, R. L., Szydlo, M., & Yung, M. (2006). Fourth-factor authentication: somebody you know (pp. 168–178). Presented at the ACM conference on computer and communications security.
- Brown, A. S., Bracken, E., Zoccoli, S., & Douglas, K. (2004). Generating and remembering passwords. *Applied Cognitive Psychology*, 18(6), 641–651.
- Buhrmester, M. D., Talafar, S., & Gosling, S. D. (2018). An Evaluation of Amazon's Mechanical Turk, Its Rapid Rise, and Its Effective Use. *Perspectives on Psychological Science*, 13(2), 149–154.
- Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 6(1), 3–5.
- Burr, W. E., Dodson, D. F., & Polk, W. T. (2004). *Electronic authentication guideline*. National Institute of Standards and Technology (NIST).
- Campbell, J., Ma, W., & Kleeman, D. (2011). Impact of restrictive composition policy on user password choices. *Behaviour & IT*, 30(3), 379–388.
- Carlton, S., Taylor, J., & Wyszynski, J. (1988). Alternate authentication mechanisms

- (pp. 333–338). Presented at the Proceedings of the Eleventh National Computer Security Conference.
- Carnavalet, X. D. C. D., & Mannan, M. (2015). A Large-Scale Evaluation of High-Impact Password Strength Meters. *ACM Trans. Inf. Syst. Secur.*, *18*(1), 1:1—1:32.
- Carroll, J., & Mack, R. (1984). Learning to use a word processor: By doing, by thinking, and by knowing. *Human Factors in Computer Systems*.
- Carstens, D. S., Malone, L. C., & McCauley-Bell, P. (2006). Applying chunking theory in organizational password guidelines. *Journal of Information, Information Technology, and Organizations*, *1*, 97–113.
- Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon’s MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior*, *29*(6), 2156–2160.
- Castelluccia, C., Dürmuth, M., & Perito, D. (2012). Adaptive Password-Strength Meters from Markov Models. Presented at the NDSS.
- Chong, M. K., & Marsden, G. (2009). Exploring the use of discrete gestures for authentication. In *Human-Computer Interaction—INTERACT 2009* (pp. 205–213). Springer.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46.
- Conlan, R. M., & Tarasewich, P. (2006). Improving interface designs to help users choose better passwords (p. 652). Presented at the CHI ‘06 extended abstracts on Human factors in computing systems - CHI EA ’06, New York, New York, USA: ACM Press.
- Coventry, L. (2005). Usable biometrics. In *Security and usability: designing secure systems that people can use*. (pp. 422–430).
- Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(01), 87–114.
- Das, A., Bonneau, J., Caesar, M., Borisov, N., & Wang, X. (2014). The Tangled Web of Password Reuse.
- De Alvare, A. M., Schultz, E. E., & Ne, J. R. (1988). A framework for password selection.
- Denning, D. E., & MacDoran, P. F. (1996). Location-based authentication: Grounding cyberspace for better security. *Computer Fraud and Security*, *1996*, 12–16.
- Dhamija, R., & Perrig, A. (2000). Deja vu : a user study using images for authentication. Presented at the 9th USENIX Security Symposium, Denver, Colorado, USA: USENIX Association.
- Dumas, J. S., & Redish, J. (1999). A practical guide to usability testing. Intellect Books.
- Egelman, S., Sotirakopoulos, A., Muslukhov, I., Beznosov, K., & Herley, C. (2013). Does my password go up to eleven?: the impact of password meters on password selection (pp. 2379–2388). Presented at the the SIGCHI Conference, New York, New York, USA: ACM.
- Fahl, S., Harbach, M., Acar, Y., & Smith, M. (2013). On the Ecological Validity of a Password Study (pp. 13:1—13:13). Presented at the Proceedings of the Ninth

- Symposium on Usable Privacy and Security, New York, NY, USA: ACM.
- Feldmeier, D. C., & Karn, P. R. (1990). Unix password security-ten years later (pp. 44–63). Presented at the Advances in Cryptology—CRYPTO'89 Proceedings, Springer.
- Federal Information Processing Standards (FIPS). (1985). *Password Usage*.
- Florencio, D., & Herley, C. (2007). A large-scale study of web password habits (p. 657). Presented at the Proceedings of the 16th international conference on World Wide Web - WWW '07, New York, New York, USA: ACM Press.
- Forget, A., Chiasson, S., van Oorschot, P. C., & Biddle, R. (2008). Improving text passwords through persuasion (pp. 1–12). Presented at the Proceedings of the 4th symposium on Usable privacy and security, ACM.
- Furnell, S. (2007). An assessment of website password practices. *Computers and Security*, 26(7), 445–451.
- Furnell, S. (2011). Assessing password guidance and enforcement on leading websites. *Computer Fraud and Security*, 2011(12), 10–18.
- Furnell, S., & Bär, N. (2013). Essential Lessons Still Not Learned? Examining the Password Practices of End-Users and Service Providers. In *Human Aspects of Information Security, Privacy, and Trust* (Vol. 8030, pp. 217–225). Berlin, Heidelberg: Springer Berlin Heidelberg.
- Furnell, S., & Esmael, R. (2017). Evaluating the effect of guidance and feedback upon password compliance. *Computer Fraud and Security*, 2017(1), 5–10.
- Garfinkel, S., & Lipford, H. R. (2014). Usable Security: History, Themes, and Challenges. *Synthesis Lectures on Information Security, Privacy, and Trust*, 5(2), 1–124.
- Gaw, S., & Felten, E. W. (2006). Password management strategies for online accounts (p. 44). Presented at the Proceedings of the second symposium on Usable privacy and security - SOUPS '06, New York, New York, USA: ACM Press.
- Grampp, F. T., & Morris, R. H. (1984). The UNIX system: UNIX operating system security. *AT&T Bell Laboratories Technical Journal*, 63(8), 1649–1672.
- Grassi, P. A., Fenton, J. L., Newton, E. M., Perlner, R. A., Regenscheid, A. R., Burr, W. E., et al. (2017). *Digital identity guidelines: authentication and lifecycle management*. Gaithersburg, MD: National Institute of Standards and Technology.
- Grawemeyer, B., & Johnson, H. (2011). Using and managing multiple passwords: A week to a view. *Interacting with Computers*, 23(3), 256–267.
- Gredler, C. (2012). White Paper – Mitigating the Risk of Poor Password Practices. Retrieved from <http://www.csid.com/resources/white-papers/white-paper-mitigating-the-risk-of-poor-password-practices/>
- Groß, T., Coopamootoo, K. P. L., & Al-Jabri, A. (2016). Effect of Cognitive Effort on Password Choice. Presented at the Twelfth Symposium on Usable Privacy and Security, Denver, CO.
- Herley, C., & Van Oorschot, P. (2012). A research agenda acknowledging the persistence of passwords. *Security & Privacy, IEEE*, 10(1), 28–36.
- Herley, C., van Oorschot, P. C., & Patrick, A. S. (2009). Passwords: If we're so smart, why are we still using them? *Financial Cryptography and Data Security*,

- 230–237.
- Hertzum, M., & Jacobsen, N. E. (2001). The Evaluator Effect - A Chilling Fact About Usability Evaluation Methods. *Int. J. Hum. Comput. Interaction*, 13(4), 421–443.
- Hertzum, M., Jacobsen, N. E., & Molich, R. (2002). Usability inspections by groups of specialists: perceived agreement in spite of disparate observations (pp. 662–663). Presented at the CHI'02 extended abstracts on Human factors in computing systems, ACM.
- Inglesant, P. G., & Sasse, M. A. (2010). The true cost of unusable password policies (p. 383). Presented at the Proceedings of the 28th international conference on Human factors in computing systems - CHI '10, New York, New York, USA: ACM Press.
- Jaferian, P., Hawkey, K., Sotirakopoulos, A., Velez-Rojas, M., & Beznosov, K. (2014). Heuristics for Evaluating IT Security Management Tools. *Human-Computer Interaction*, 29(4), 311–350.
- Jain, A. K., Ross, A., & Prabhakar, S. (2004). An Introduction to Biometric Recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(1), 4–20.
- Jain, A., Hong, L., & Pankanti, S. (2000). Biometric identification. *Communications of the ACM*, 43(2), 90–98.
- Jeffries, R., Miller, J. R., Wharton, C., & Uyeda, K. (1991). User interface evaluation in the real world: a comparison of four techniques (pp. 119–124). Presented at the Proceedings of the SIGCHI conference on Human factors in computing systems, ACM.
- Just, M. (2005). Designing authentication systems with challenge questions. In L. F. Cranor & S. Garfinkel (Eds.), *Security and Usability: Designing Secure Systems That People Can Use* (First, pp. 143–155). Sebastopol: O'Reilly Media.
- Just, M., & Aspinall, D. (2009). Personal choice and challenge questions: a security and usability assessment (p. 8). Presented at the Proceedings of the 5th Symposium on Usable Privacy and Security, ACM.
- Kafas, K., Aljaffan, N., & Li, S. (2013). Poster: Visual Password Checker. *Symposium on Usable Privacy and Security (SOUPS)*.
- Karreman, J., Ummelen, N., & Steehouder, M. (2005). Procedural and declarative information in user instructions: What we do and don't know about these information types (pp. 328–333). Presented at the Professional Communication Conference, 2005. IPCC 2005. Proceedings. International, IEEE.
- Kelley, P. G., Komanduri, S., Mazurek, M. L., Shay, R., Vidas, T., Bauer, L., et al. (2012). Guess again (and again and again): Measuring password strength by simulating password-cracking algorithms (pp. 523–537). Presented at the Security and Privacy (SP), 2012 IEEE Symposium on, IEEE.
- Khern-am-nuai, W., Yang, W., & Li, N. (2017). Using Context-Based Password Strength Meter to Nudge Users' Password Generating Behavior: A Randomized Experiment. Presented at the Hawaii International Conference on System Sciences, Hawaii International Conference on System Sciences.
- Klein, D. V. (1990). Foiling the cracker: A survey of, and improvements to, password security (pp. 5–14). Presented at the Proceedings of the 2nd USENIX

- Security Workshop.
- Komanduri, S., Shay, R., Kelley, P. G., Mazurek, M. L., Bauer, L., Christin, N., et al. (2011). Of passwords and people: measuring the effect of password-composition policies (p. 2595). Presented at the Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11.
- Kuo, C., Romanosky, S., & Cranor, L. F. (2006). Human selection of mnemonic phrase-based passwords (pp. 67–78). Presented at the Proceedings of the second symposium on Usable privacy and security, ACM.
- Lazar, J., Feng, J. H., & Hochheiser, H. (2017). *Research Methods in Human-Computer Interaction*. (T. Green, Ed.) (Second Edition, pp. 1–562). Cambridge: Todd Green.
- Maguire, J., & Renaud, K. (2012). You only live twice or "the years we wasted caring about shoulder-surfing". *Bcs Hci*.
- Mankoff, J., Dey, A. K., Hsieh, G., Kientz, J. A., Lederer, S., & Ames, M. (2003). Heuristic evaluation of ambient displays. *Chi*, 169.
- Miller, B. (1994). Vital signs of identity [biometrics]. *IEEE Spectrum*, 31(2), 22–30.
- Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, 63(2), 81–97.
- Morris, R., & Thompson, K. (1979). Password security: a case history. *Communications of the ACM*, 22(11), 594–597.
- Nielsen, J. (1994). Heuristic evaluation. *Usability Inspection Methods*, 17(1), 25–62.
- Norman, D. (2002). *The design of everyday things*. New York: Basic books.
- Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, 45(4), 867–872.
- Paz, F., Paz, F. A., & Pow-Sang, J. A. (2015). Experimental case study of new usability heuristics (pp. 212–223). Presented at the International Conference of Design, User Experience, and Usability, Springer.
- Peer, E., Vosgerau, J., & Acquisti, A. (2013). Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods*, 46(4), 1023–1031.
- Petrie, H., & Buykx, L. (2010). Collaborative Heuristic Evaluation: improving the effectiveness of heuristic evaluation. Presented at the Proceedings of UPA 2010 International Conference. Omnipress.
- Petrie, H., & Power, C. (2012). What do users really care about?: a comparison of usability problems found by users and experts on highly interactive websites (pp. 2107–2116). Presented at the Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems, ACM.
- Pfleeger, C. P., & Pfleeger, S. L. (2006). *Security in computing*. Prentice Hall PTR.
- Pinelle, D., Wong, N., & Stach, T. (2008). Heuristic evaluation for games: usability principles for video game design. *Proceeding of the twenty-sixth annual CHI conference* (pp. 1453–1462). New York, New York, USA: ACM.
- Ponemon Institute. (2006). *Perceptions About Passwords*. Retrieved from <http://www.csoonline.com/article/2117899/identity-access/those-pesky-passwords.html>
- Proctor, R. W., Lien, M.-C., Vu, K.-P. L., Schultz, E. E., & Salvendy, G. (2002).

- Improving computer security for authentication of users: Influence of proactive password restrictions. *Behavior Research Methods, Instruments, & Computers*, 34(2), 163–169.
- Renaud, K. (2005). Evaluating Authentication Mechanisms. In L. Cranor & S. Garfinkel (Eds.), *Security and Usability* (pp. 103–128). O'Reilly Media Inc.
- Riddle, B. L., Miron, M. S., & Semo, J. A. (1989). Passwords in use in a university timesharing environment. *Computers and Security*, 8(7), 569–579.
- Sasse, M. A., Brostoff, S., & Weirich, D. (2001). Transforming the “Weakest Link” — a Human/Computer Interaction Approach to Usable and Effective Security. *BT Technology Journal*, 19(3), 122–131.
- Sauro, J. (2010). *A Practical Guide to Measuring Usability: 72 Answers to the Most Common Questions about Quantifying the Usability of Websites and Software*. Measuring Usability LLC.
- Schechter, S. E., Dhamija, R., Ozment, A., & Fischer, I. (2007). The Emperor's New Security Indicators (pp. 51–65). Presented at the 2007 IEEE Symposium on Security and Privacy (SP '07), IEEE.
- Schneier, B. (1999). Inside risks: the uses and abuses of biometrics. *Communications of the ACM*, 42, 136.
- Shay, R. (2015). *Creating Usable Policies for Stronger Passwords with MTurk*.
- Shay, R., Bauer, L., Christin, N., Cranor, L. F., Forget, A., Komanduri, S., et al. (2015). A Spoonful of Sugar?: The Impact of Guidance and Feedback on Password-Creation Behavior (pp. 2903–2912). Presented at the Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, New York, NY, USA: ACM.
- Shay, R., Komanduri, S., Durity, A. L., Huh, P. S., Mazurek, M. L., Segreti, S. M., et al. (2014). Can long passwords be secure and usable? *Chi*, 2927–2936.
- Shay, R., Komanduri, S., Kelley, P. G., Leon, P. G., Mazurek, M. L., Bauer, L., et al. (2010). Encountering Stronger Password Requirements: User Attitudes and Behaviors (p. 1). Presented at the Proceedings of the Sixth Symposium on Usable Privacy and Security - SOUPS '10, New York, New York, USA: ACM Press.
- Shneiderman, B., Plaisant, C., Cohen, M. S., Jacobs, S., & Elmqvist, N. (2018). *Designing the user interface*. Pearson.
- Smith, E. E., & Goodman, L. (1984). Understanding Written Instructions: The Role of an Explanatory Schema. *Cognition and Instruction*, 1(4), 359–396.
- Smith, R. E. (2001). *Authentication: From Passwords to Public Keys* (First edit). Addison Wesley.
- Spafford, E. H. (1989). The Internet worm program: An analysis. *ACM SIGCOMM Computer Communication Review*, 19(1), 17–57.
- Sweller, J., & Chandler, P. (1991). Evidence for cognitive load theory. *Cognition and Instruction*, 8(4), 351–362.
- Tam, L., Glassman, M., & Vandenwauver, M. (2010). The psychology of password management: a tradeoff between security and convenience. *Behaviour & Information Technology*, 29(3), 233–244.
- Tan, W., Hsu, J., & Pinn, F. (2001). *Method and system for token-based authentication*. Google Patents.

- Technologies, S. (2003, June). Password usage survey. Retrieved from http://www.safenet-inc.com/solutions/password_survey.asp
- Ummelen, N. (1997). Declarative information in software manuals (pp. 283–296). Presented at the the 15th annual international conference, New York, New York, USA: ACM Press.
- Ur, B., Kelley, P. G., Komanduri, S., Lee, J., Maass, M., Mazurek, M. L., et al. (2012). How Does Your Password Measure Up? The Effect of Strength Meters on Password Creation. (pp. 65–80). Presented at the USENIX Security Symposium.
- Ur, B., Segreti, S. M., Bauer, L., Christin, N., Cranor, L. F., Komanduri, S., et al. (2015). Measuring Real-World Accuracies and Biases in Modeling Password Guessability. *USENIX Security Symposium*.
- US Department of Defense. (1985). *Password Management Guidelines*. CSC-STD-002-85.
- Van den Haak, M. J., & de Jong, M. D. (2003). Exploring two methods of usability testing: concurrent versus retrospective think-aloud protocols (pp. 3–pp). Presented at the Professional Communication Conference, 2003. IPCC 2003. Proceedings. IEEE International.
- Vance, A., Eargle, D., Ouimet, K., & Straub, D. W. (2013). Enhancing Password Security through Interactive Fear Appeals - A Web-Based Field Experiment. *Hicss*, 2988–2997.
- Vu, K.-P. L., Proctor, R. W., Bhargav-Spantzel, A., Tai, B.-L. B., Cook, J., & Eugene Schultz, E. (2007). Improving password security and memorability to protect personal and organizational information. *International Journal of Human-Computer Studies*, 65(8), 744–757.
- Wharton, C., Bradford, J., Jeffries, R., & Franzke, M. (1992). Applying cognitive walkthroughs to more complex user interfaces - experiences, issues, and recommendations. *Chi*.
- Wiedenbeck, S., Waters, J., Birget, J.-C., Brodskiy, A., & Memon, N. (2005). Authentication using graphical passwords: effects of tolerance and image choice (pp. 1–12). Presented at the Proceedings of the 2005 symposium on Usable privacy and security, ACM.
- Wiedenbeck, S., Waters, J., Sobrado, L., & Birget, J.-C. (2006). Design and evaluation of a shoulder-surfing resistant graphical password scheme (pp. 177–184). Presented at the Proceedings of the working conference on Advanced visual interfaces, ACM.
- Yan, Jeff, Blackwell, A., Anderson, R., & Grant, A. (2004). Password memorability and security: Empirical results. *Security & Privacy, IEEE*, 2, 25–31.
- Yan, Jianxin, Blackwell, A., Anderson, R., & Grant, A. (2000). *The memorability and security of passwords – some empirical results*. Cambridge, United Kingdom: University of Cambridge Computer Laboratory.

Appendix A

Informed Consent

A.1 Consent Form (Online Studies)

Study on instructions for creating passwords

Dear participant,

Thank you for offering to participate in this study. The study is part of my PhD study at the University of York. I am investigating how different kinds of instructions help or hinder people when they create passwords.

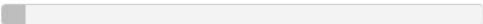
During the study you will be asked to rate a number of different instructions you might see when you are creating a password for an online system. These instructions can be presented at two different stages during the password creation process:
Stage 1: Before you start creating a password, so when you open the page with the field for entering the password;
Stage 2: As you enter your proposed password.

In this study I will not ask you to reveal any of your passwords or create any passwords. I'm also not testing you in any way, my research aims to create clearer, more effective password instructions.

All the information you provide will be completely confidential. The only people who will see it will be myself and my supervisor, Professor Helen Petrie. Any responses that you contribute to this study will not be attributed to you, nor will you be individually identified in any results.

You may choose to withdraw from the study at any time. If you have any questions about this study, please contact me, Saja Althubaiti (saaa505@york.ac.uk) or my supervisor Professor Helen Petrie (helen.petrie@york.ac.uk).

By clicking 'Next' you are giving your informed consent that you agree and understand the information above.

1 / 20  5%

Next

A.2 Consent Form (Lab Studies)

User-based Evaluation on Password Creation Systems

Thank you for participating in this study. This study is part of my PhD work at the University of York in Computer Science Department. As part of my work, I am working on how to improve the effectiveness of the current password creation systems.

Taking part in this study will not compromise your passwords or computer security in any way. During the study you will be asked to evaluate six password creation systems by trying a set of passwords. The set of passwords is not used for real systems. After this you will be asked to answer some demographic information about yourself. All the information you provide will be completely confidential.

The only people who will see it will be myself and my supervisor, Professor Helen Petrie.

Any responses that you contribute to this evaluation will not be attributed to you, nor will you be individually identified in any results. Any results will be reported in an aggregate form with the responses from other participants. You may choose to withdraw the experiment at any time.

If you have any questions about this study or anything else you would like to raise, then please contact me at saaa505@york.ac.uk or my supervisor Helen Petrie at helen.petrie@york.ac.uk

Please sign below that you agree to take part in the study under the conditions laid out above. This will indicate that you have read and understood the above and that I will be obliged to treat your data as described.

Name:

Signature:

Date:

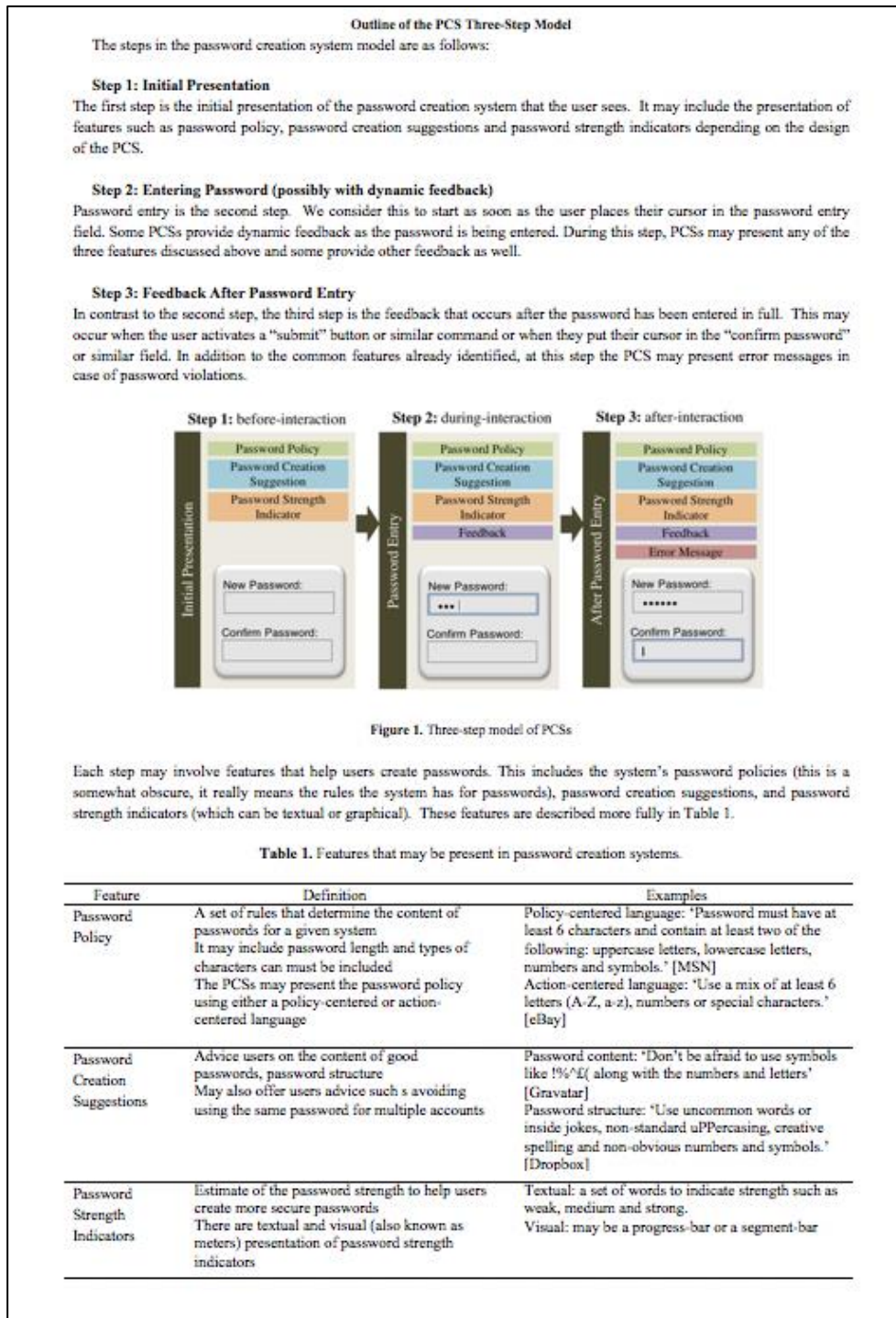
Appendix B

Usability Evaluations

B.1 List of Provided Passwords (Studies 2 and 3)

Password	Length	Composition of character classes			
		<i>Digits</i>	<i>Uppercase letters</i>	<i>Lowercase letters</i>	<i>Symbols</i>
12	2	2	*	*	*
123	3	3	*	*	*
Mm5	3	1	1	1	*
1234	4	4	*	*	*
12345	5	5	*	*	*
Mmouse	6	*	1	5	*
Mm1234@	7	4	1	1	1
12345678	8	8	*	*	*
1234uuuM	8	4	1	3	*
password	8	*	*	8	*
Password	8	*	1	7	*
mM123456	8	6	1	1	*
mM1abcde	8	1	1	6	*
M(mouse)	8	*	1	5	2
Mmouse123	9	3	1	5	*
Mmouse123@	10	3	1	5	1
Mmouse1234!	11	4	1	5	1
MickeyMouse12	13	2	2	9	*
MickeyMouse123	14	3	2	9	*
Mickeymouse123456789	20	9	1	10	*
Mickeymouse123456789(?)	23	9	1	10	3

B.2 Outline of the PCS Three-Step Model (Study 2)



B.3 Demographic Questionnaire (Study 3)

Questionnaire for User-based Evaluation of Password Creation Systems

1. Sex (M/F):
2. Age:
3. Is your native language English? (a) Yes (b) No
If No, what is your native language and how long have you been speaking English? _____
4. Current occupation:
(a) Student (b) Other, please specify _____
5. On average, how often do you use the Internet in a day?
(a) Never (b) Rarely (c) Sometime (d) Often (e) Always
6. On average, how often do you use any type of computer (e.g. Desktop, Laptop, Tablet) in a day?
(a) Never (b) Rarely (c) Sometime (d) Often (e) Always
7. Have you created an account in the last month for the following websites:

1. Apple	(a) Yes	(b) No	(c) Don't think so
2. Wikipedia	(a) Yes	(b) No	(c) Don't think so
3. Netflix	(a) Yes	(b) No	(c) Don't think so
4. WordPress	(a) Yes	(b) No	(c) Don't think so
5. Dailymail/ Mailonline	(a) Yes	(b) No	(c) Don't think so
6. Stackoverflow	(a) Yes	(b) No	(c) Don't think so

Email (for Amazon voucher only):

This information will be destroyed after you are sent the voucher.

Appendix C

User Instructions

C.1 Difference Between Non-MTurkers and MTurkers

A Mann-Whitney U Test Independent-Samples was used to examine the difference between non-MTurkers and MTurkers (referred to as Data Source in the table below). The tests from 1 to 72 are for Group 2 and from 73 to 157 are for Group 3.

	Null Hypothesis	Sig.
1	The distribution of SS1_Cond1.1_DeclPasswordAbstract_Helpful is the same across categories of DataSource.	.706 ^a
2	The distribution of SS1_Cond1.1_DeclPasswordAbstract_Clear is the same across categories of DataSource.	.179 ^a
3	The distribution of SS1_Cond1.1_DeclPasswordAbstract_DetailsRecode is the same across categories of DataSource.	.706 ^a
4	The distribution of SS1_Cond1.1_DeclPasswordAbstract_Confidence is the same across categories of DataSource.	.179 ^a
5	The distribution of SS1_Cond1.2_DeclPasswordConcrete_Helpful is the same across categories of DataSource.	.732 ^a
6	The distribution of SS1_Cond1.2_DeclPasswordConcrete_Clear is the same across categories of DataSource.	.167 ^a
7	The distribution of SS1_Cond1.2_DeclPasswordConcrete_DetailsRecode is the same across categories of DataSource.	.286 ^a
8	The distribution of SS1_Cond1.2_DeclPasswordConcrete_Confidence is the same across categories of DataSource.	.167 ^a
9	The distribution of SS1_Cond2.3_DeclActionAbstract_Helpful is the same across categories of DataSource.	.973 ^a
10	The distribution of SS1_Cond2.3_DeclActionAbstract_Clear is the same across categories of DataSource.	.784 ^a
11	The distribution of SS1_Cond2.3_DeclActionAbstract_DetailsRecode is the same across categories of DataSource.	.515 ^a
12	The distribution of SS1_Cond2.3_DeclActionAbstract_Confidence is the same across categories of DataSource.	.758 ^a
13	The distribution of SS1_Cond2.4_DeclActionConcrete_Helpful is the same across categories of DataSource.	.515 ^a
14	The distribution of SS1_Cond2.4_DeclActionConcrete_Clear is the same across categories of DataSource.	.286 ^a
15	The distribution of SS1_Cond2.4_DeclActionConcrete_DetailsRecode is the same across categories of DataSource.	.891 ^a
16	The distribution of SS1_Cond2.4_DeclActionConcrete_Confidence is the same across categories of DataSource.	.515 ^a
17	The distribution of SS1_Cond3.5_ImpPositiveAbstractHowWhy_Helpful is the same across categories of DataSource.	.732 ^a

18	The distribution of SS1_Cond3.5_ImpPositiveAbstractHowWhy_Clear is the same across categories of DataSource.	.607 ^a
19	The distribution of SS1_Cond3.5_ImpPositiveAbstractHowWhy_DetailsRecode is the same across categories of DataSource.	.515 ^a
20	The distribution of SS1_Cond3.5_ImpPositiveAbstractHowWhy_Confidence is the same across categories of DataSource.	.128 ^a
21	The distribution of SS1_Cond3.6_ImpPositiveConcreteHowWhy_Helpful is the same across categories of DataSource.	.656 ^a
22	The distribution of SS1_Cond3.6_ImpPositiveConcreteHowWhy_Clear is the same across categories of DataSource.	.047 ^a
23	The distribution of SS1_Cond3.6_ImpPositiveConcreteHowWhy_DetailsRecode is the same across categories of DataSource.	.286 ^a
24	The distribution of SS1_Cond3.6_ImpPositiveConcreteHowWhy_Confidence is the same across categories of DataSource.	.811 ^a
25	The distribution of SS1_Cond3.7_ImpPositiveAbstractHow_Helpful is the same across categories of DataSource.	.372 ^a
26	The distribution of SS1_Cond3.7_ImpPositiveAbstractHow_Clear is the same across categories of DataSource.	.973 ^a
27	The distribution of SS1_Cond3.7_ImpPositiveAbstractHow_DetailsRecode is the same across categories of DataSource.	.837 ^a
28	The distribution of SS1_Cond3.7_ImpPositiveAbstractHow_Confidence is the same across categories of DataSource.	.372 ^a
29	The distribution of SS1_Cond4.8_ImpPositivePolite_Helpful is the same across categories of DataSource.	.656 ^a
30	The distribution of SS1_Cond4.8_ImpPositivePolite_Clear is the same across categories of DataSource.	.410 ^a
31	The distribution of SS1_Cond4.8_ImpPositivePolite_DetailsRecode is the same across categories of DataSource.	.632 ^a
32	The distribution of SS1_Cond4.8_ImpPositivePolite_Confidence is the same across categories of DataSource.	.215 ^a
33	The distribution of SS1_Cond5.9_ImpNegative_Helpful is the same across categories of DataSource.	.681 ^a
34	The distribution of SS1_Cond5.9_ImpNegative_Clear is the same across categories of DataSource.	1.000 ^a
35	The distribution of SS1_Cond5.9_ImpNegative_DetailsRecode is the same across categories of DataSource.	.758 ^a
36	The distribution of SS1_Cond5.9_ImpNegative_Confidence is the same across categories of DataSource.	.515 ^a
37	The distribution of SS2_Cond1.1_DeclGeneral_Helpful is the same across categories of DataSource.	.190 ^a
38	The distribution of SS2_Cond1.1_DeclGeneral_Clear is the same across categories of DataSource.	.202 ^a
39	The distribution of SS2_Cond1.1_DeclGeneral_DetailsRecode is the same across categories of DataSource.	.632 ^a
40	The distribution of SS2_Cond1.1_DeclGeneral_Confidence is the same across categories of DataSource.	.471 ^a
41	The distribution of SS2_Cond2.2_DeclSpecificAbstract_Helpful is the same across categories of DataSource.	.430 ^a
42	The distribution of SS2_Cond2.2_DeclSpecificAbstract_Clear is the same across categories of DataSource.	.471 ^a
43	The distribution of SS2_Cond2.2_DeclSpecificAbstract_DetailsRecode is the same across categories of DataSource.	.607 ^a
44	The distribution of SS2_Cond2.2_DeclSpecificAbstract_Confidence is the same across categories of DataSource.	.202 ^a
45	The distribution of SS2_Cond2.3_DeclSpecificConcrete_Helpful is the same across categories of DataSource.	.319 ^a
46	The distribution of SS2_Cond2.3_DeclSpecificConcrete_Clear is the same across categories of DataSource.	.286 ^a
47	The distribution of SS2_Cond2.3_DeclSpecificConcrete_DetailsRecode is the same across categories of DataSource.	.537 ^a

48	The distribution of SS2_Cond2.3_DeclSpecificConcrete_Confidence is the same across categories of DataSource.	.784 ^a
49	The distribution of SS2_Cond3.4_ImpGeneral_Helpful is the same across categories of DataSource.	.973 ^a
50	The distribution of SS2_Cond3.4_ImpGeneral_Clear is the same across categories of DataSource.	.451 ^a
51	The distribution of SS2_Cond3.4_ImpGeneral_DetailsRecode is the same across categories of DataSource.	.732 ^a
52	The distribution of SS2_Cond3.4_ImpGeneral_Confidence is the same across categories of DataSource.	.607 ^a
53	The distribution of SS2_Cond4.5_ImpGeneralPolite_Helpful is the same across categories of DataSource.	1.000 ^a
54	The distribution of SS2_Cond4.5_ImpGeneralPolite_Clear is the same across categories of DataSource.	.784 ^a
55	The distribution of SS2_Cond4.5_ImpGeneralPolite_DetailsRecode is the same across categories of DataSource.	.945 ^a
56	The distribution of SS2_Cond4.5_ImpGeneralPolite_Confidence is the same across categories of DataSource.	.430 ^a
57	The distribution of SS2_Cond5.6_ImpSpecificAbstract_Helpful is the same across categories of DataSource.	.036 ^a
58	The distribution of SS2_Cond5.6_ImpSpecificAbstract_Clear is the same across categories of DataSource.	.056 ^a
59	The distribution of SS2_Cond5.6_ImpSpecificAbstract_DetailsRecode is the same across categories of DataSource.	.228 ^a
60	The distribution of SS2_Cond5.6_ImpSpecificAbstract_Confidence is the same across categories of DataSource.	.003 ^a
61	The distribution of SS2_Cond5.7_ImpSpecificConcrete_Helpful is the same across categories of DataSource.	.167 ^a
62	The distribution of SS2_Cond5.7_ImpSpecificConcrete_Clear is the same across categories of DataSource.	.537 ^a
63	The distribution of SS2_Cond5.7_ImpSpecificConcrete_DetailsRecode is the same across categories of DataSource.	.681 ^a
64	The distribution of SS2_Cond5.7_ImpSpecificConcrete_Confidence is the same across categories of DataSource.	.837 ^a
65	The distribution of SS2_Cond6.8_ImpSpecificPolite_Helpful is the same across categories of DataSource.	.104 ^a
66	The distribution of SS2_Cond6.8_ImpSpecificPolite_Clear is the same across categories of DataSource.	.096 ^a
67	The distribution of SS2_Cond6.8_ImpSpecificPolite_DetailsRecode is the same across categories of DataSource.	.302 ^a
68	The distribution of SS2_Cond6.8_ImpSpecificPolite_Confidence is the same across categories of DataSource.	.043 ^a
69	The distribution of SS2_Cond7.9_ImperativeSpecificNegative_Helpful is the same across categories of DataSource.	.607 ^a
70	The distribution of SS2_Cond7.9_ImperativeSpecificNegative_Clear is the same across categories of DataSource.	.837 ^a
71	The distribution of SS2_Cond7.9_ImperativeSpecificNegative_DetailsRecode is the same across categories of DataSource.	1.000 ^a
72	The distribution of SS2_Cond7.9_ImperativeSpecificNegative_Confidence is the same across categories of DataSource.	.607 ^a
73	The distribution of ES2_Cond1_DeclGeneralPositive_Helpful is the same across categories of DataSource.	.432
74	The distribution of ES2_Cond1_DeclGeneralPositive_Clear is the same across categories of DataSource.	.012
75	The distribution of ES2_Cond1_DeclGeneralPositive_Details is the same across categories of DataSource.	.037
76	The distribution of ES2_Cond1_DeclGeneralPositive_DetailsRecode is the same across categories of DataSource.	.321
77	The distribution of ES2_Cond1_DeclGeneralPositive_Confidence is the same across categories of DataSource.	.002

77	The distribution of ES2_Cond2_DeclGeneralNegative_Helpful is the same across categories of DataSource.	.573
78	The distribution of ES2_Cond2_DeclGeneralNegative_Clear is the same across categories of DataSource.	.168
79	The distribution of ES2_Cond2_DeclGeneralNegative_Details is the same across categories of DataSource.	.678
80	The distribution of ES2_Cond2_DeclGeneralNegative_DetailsRecode is the same across categories of DataSource.	.916
81	The distribution of ES2_Cond2_DeclGeneralNegative_Confidence is the same across categories of DataSource.	.099
82	The distribution of ES2_Cond3_DeclSpecificNegative_Helpful is the same across categories of DataSource.	.751
83	The distribution of ES2_Cond3_DeclSpecificNegative_Clear is the same across categories of DataSource.	.715
84	The distribution of ES2_Cond3_DeclSpecificNegative_Details is the same across categories of DataSource.	.302
85	The distribution of ES2_Cond3_DeclSpecificNegative_DetailsRecode is the same across categories of DataSource.	.733
86	The distribution of ES2_Cond3_DeclSpecificNegative_Confidence is the same across categories of DataSource.	.089
87	The distribution of PS3_Cond1_DeclarativePositive_Helpful is the same across categories of DataSource.	.603
88	The distribution of PS3_Cond1_DeclarativePositive_Clear is the same across categories of DataSource.	.207
89	The distribution of PS3_Cond1_DeclarativePositive_Details is the same across categories of DataSource.	.875
90	The distribution of PS3_Cond1_DeclarativePositive_DetailsRecode is the same across categories of DataSource.	.644
91	The distribution of PS3_Cond1_DeclarativePositive_Confidence is the same across categories of DataSource.	.210
92	The distribution of PS3_Cond2_ModalPositive_Helpful is the same across categories of DataSource.	.046
93	The distribution of PS3_Cond2_ModalPositive_Clear is the same across categories of DataSource.	.047
94	The distribution of PS3_Cond2_ModalPositive_Details is the same across categories of DataSource.	.412
95	The distribution of PS3_Cond2_ModalPositive_DetailsRecode is the same across categories of DataSource.	.748
96	The distribution of PS3_Cond2_ModalPositive_Confidence is the same across categories of DataSource.	.185
97	The distribution of PS3_Cond3_ModalNegative_Helpful is the same across categories of DataSource.	.011
98	The distribution of PS3_Cond3_ModalNegative_Clear is the same across categories of DataSource.	.001
99	The distribution of PS3_Cond3_ModalNegative_Details is the same across categories of DataSource.	.188
100	The distribution of PS3_Cond3_ModalNegative_DetailsRecode is the same across categories of DataSource.	.012
101	The distribution of PS3_Cond3_ModalNegative_Confidence is the same across categories of DataSource.	.001
102	The distribution of PS3_Cond4_ImperativePositive_Helpful is the same across categories of DataSource.	.192
103	The distribution of PS3_Cond4_ImperativePositive_Clear is the same across categories of DataSource.	.201
104	The distribution of PS3_Cond4_ImperativePositive_Details is the same across categories of DataSource.	.133
105	The distribution of PS3_Cond4_ImperativePositive_DetailsRecode is the same across categories of DataSource.	.950
106	The distribution of PS3_Cond4_ImperativePositive_Confidence is the same across categories of DataSource.	.471

107	The distribution of PS3_Cond5_ImperativePositivePolite_Helpful is the same across categories of DataSource.	.198
108	The distribution of PS3_Cond5_ImperativePositivePolite_Clear is the same across categories of DataSource.	.371
109	The distribution of PS3_Cond5_ImperativePositivePolite_Details is the same across categories of DataSource.	.091
110	The distribution of PS3_Cond5_ImperativePositivePolite_DetailsRecode is the same across categories of DataSource.	.710
111	The distribution of PS3_Cond5_ImperativePositivePolite_Confidence is the same across categories of DataSource.	.397
112	The distribution of PS3_Cond6_ImperativeNegative_Helpful is the same across categories of DataSource.	.033
113	The distribution of PS3_Cond6_ImperativeNegative_Clear is the same across categories of DataSource.	.023
114	The distribution of PS3_Cond6_ImperativeNegative_Details is the same across categories of DataSource.	.193
115	The distribution of PS3_Cond6_ImperativeNegative_DetailsRecode is the same across categories of DataSource.	.050
116	The distribution of PS3_Cond6_ImperativeNegative_Confidence is the same across categories of DataSource.	.034
117	The distribution of SS3_Cond1_DeclSpecificConcrete_Helpful is the same across categories of DataSource.	.568
118	The distribution of SS3_Cond1_DeclSpecificConcrete_Clear is the same across categories of DataSource.	.116
119	The distribution of SS3_Cond1_DeclSpecificConcrete_Details is the same across categories of DataSource.	.191
120	The distribution of SS3_Cond1_DeclSpecificConcrete_DetailsRecode is the same across categories of DataSource.	.637
121	The distribution of SS3_Cond1_DeclSpecificConcrete_Confidence is the same across categories of DataSource.	.077
122	The distribution of SS3_Cond2_ImperativeGeneral_Helpful is the same across categories of DataSource.	.366
123	The distribution of SS3_Cond2_ImperativeGeneral_Clear is the same across categories of DataSource.	.147
124	The distribution of SS3_Cond2_ImperativeGeneral_Details is the same across categories of DataSource.	.308
125	The distribution of SS3_Cond2_ImperativeGeneral_DetailsRecode is the same across categories of DataSource.	.450
126	The distribution of SS3_Cond2_ImperativeGeneral_Confidence is the same across categories of DataSource.	.109
127	The distribution of SS3_Cond3_ImperativeGeneralPolite_Helpful is the same across categories of DataSource.	.424
128	The distribution of SS3_Cond3_ImperativeGeneralPolite_Clear is the same across categories of DataSource.	.436
129	The distribution of SS3_Cond3_ImperativeGeneralPolite_Details is the same across categories of DataSource.	.368
130	The distribution of SS3_Cond3_ImperativeGeneralPolite_DetailsRecode is the same across categories of DataSource.	.720
131	The distribution of SS3_Cond3_ImperativeGeneralPolite_Confidence is the same across categories of DataSource.	.113
132	The distribution of SS3_Cond4_ImperativeSpecific_Helpful is the same across categories of DataSource.	.520
133	The distribution of SS3_Cond4_ImperativeSpecific_Clear is the same across categories of DataSource.	.266
134	The distribution of SS3_Cond4_ImperativeSpecific_Details is the same across categories of DataSource.	.821
135	The distribution of SS3_Cond4_ImperativeSpecific_DetailsRecode is the same across categories of DataSource.	.103

132	The distribution of SS3_Cond4_ImperativeSpecific_Confidence is the same across categories of DataSource.	.669
133	The distribution of ES3_Cond1_DeclarativeGeneralNegativePolite_Helpful is the same across categories of DataSource.	.030
134	The distribution of ES3_Cond1_DeclarativeGeneralNegativePolite_Clear is the same across categories of DataSource.	.010
135	The distribution of ES3_Cond1_DeclarativeGeneralNegativePolite_Details is the same across categories of DataSource.	.044
136	The distribution of ES3_Cond1_DeclarativeGeneralNegativePolite_DetailsRecode is the same across categories of DataSource.	.049
137	The distribution of ES3_Cond1_DeclarativeGeneralNegativePolite_Confidence is the same across categories of DataSource.	.034
138	The distribution of ES3_Cond2_DeclarativeGeneralNegative_Helpful is the same across categories of DataSource.	.457
139	The distribution of ES3_Cond2_DeclarativeGeneralNegative_Clear is the same across categories of DataSource.	.857
140	The distribution of ES3_Cond2_DeclarativeGeneralNegative_Details is the same across categories of DataSource.	.770
141	The distribution of ES3_Cond2_DeclarativeGeneralNegative_DetailsRecode is the same across categories of DataSource.	.290
142	The distribution of ES3_Cond2_DeclarativeGeneralNegative_Confidence is the same across categories of DataSource.	.524
143	The distribution of ES3_Cond3_DeclarativeGeneralPositive_Helpful is the same across categories of DataSource.	.329
144	The distribution of ES3_Cond3_DeclarativeGeneralPositive_Clear is the same across categories of DataSource.	.091
145	The distribution of ES3_Cond3_DeclarativeGeneralPositive_Details is the same across categories of DataSource.	.069
146	The distribution of ES3_Cond3_DeclarativeGeneralPositive_DetailsRecode is the same across categories of DataSource.	.099
147	The distribution of ES3_Cond3_DeclarativeGeneralPositive_Confidence is the same across categories of DataSource.	.300
148	The distribution of ES3_Cond4_DeclarativeSpecificNegative_Helpful is the same across categories of DataSource.	.269
149	The distribution of ES3_Cond4_DeclarativeSpecificNegative_Clear is the same across categories of DataSource.	.117
150	The distribution of ES3_Cond4_DeclarativeSpecificNegative_Details is the same across categories of DataSource.	.012
151	The distribution of ES3_Cond4_DeclarativeSpecificNegative_DetailsRecode is the same across categories of DataSource.	.206
152	The distribution of ES3_Cond4_DeclarativeSpecificNegative_Confidence is the same across categories of DataSource.	.273
153	The distribution of ES3_Cond5_PhraseOnlyGeneralNegative_Helpful is the same across categories of DataSource.	.015
154	The distribution of ES3_Cond5_PhraseOnlyGeneralNegative_Clear is the same across categories of DataSource.	.005
155	The distribution of ES3_Cond5_PhraseOnlyGeneralNegative_Details is the same across categories of DataSource.	.089
156	The distribution of ES3_Cond5_PhraseOnlyGeneralNegative_DetailsRecode is the same across categories of DataSource.	.173
157	The distribution of ES3_Cond5_PhraseOnlyGeneralNegative_Confidence is the same across categories of DataSource.	.010

C.2 Full Set of User Instruction

Group 1		
<i>The requirement of the system in this case (Policy in step 1) will be 6 characters and 1 numeral</i>		
1	Statement 1	<i>Declarative -> declarative -> password-oriented -> specific -> positive -> no-polite</i> 1. The password needs to have at least six characters and at least one numeral
2	Statement 2	<i>Declarative -> phrase only -> password-oriented -> specific -> positive -> no-polite</i> 2. at least six characters; at least one numeral
3	Statement 3	<i>Procedural -> imperative -> action-oriented -> specific -> positive -> no-polite</i> 3. Use at least six characters and at least one numeral
4	Statement 4	<i>Procedural -> imperative -> action-oriented -> specific -> positive -> yes-polite</i> 4. Please use at least six characters and at least one numeral
5	Statement 5	<i>Procedural -> imperative -> action-oriented -> specific -> negative -> no-polite</i> 5. Do not use less than six characters but use at least one numeral
<i>The requirement of the system in this case (Policy in step 2) will be only lowercase letters</i>		
6	Statement 1	<i>Declarative -> declarative -> password-oriented -> specific -> positive -> no-polite</i> 1. The password needs to have only lowercase letters
7	Statement 2	<i>Declarative -> declarative -> password-oriented -> specific -> negative -> no-polite</i> 2. The password needs to not contain uppercase letters
8	Statement 3	<i>Declarative -> modal -> password-oriented -> specific -> positive -> no-polite</i> 3. The password must have only lowercase letters
9	Statement 4	<i>Declarative -> modal -> password-oriented -> specific -> negative -> no-polite</i> 4. The password must not contain uppercase letters
10	Statement 5	<i>Declarative -> phrase only -> password-oriented -> specific -> positive -> no-polite</i> 5. Only lowercase letters
11	Statement 6	<i>Declarative -> phrase only -> password-oriented -> specific -> negative -> no-polite</i> 6. No uppercase letters
12	Statement 7	<i>Procedural -> imperative -> action-oriented -> specific -> positive -> no-polite</i> 7. Use only lowercase letters
13	Statement 8	<i>Procedural -> imperative -> action-oriented -> specific -> positive -> yes-polite</i> 8. Please use only lowercase letters
14	Statement 9	<i>Procedural -> imperative -> action-oriented -> specific -> negative -> no-polite</i> 9. Do not use uppercase letters
Group 2		
<i>The suggestion in this case (suggestion in step 1) will be (1) abstract: uncommon words. (2) concrete: have both letters and numbers.</i>		
15 16	Statement 1&2	<i>Declarative -> declarative -> password-oriented -> specific -> positive -> no-polite (HW)</i> 1. Good passwords have uncommon words (abstract) 2. Good passwords have both letters and numbers (concrete)
17 18	Statement 3&4	<i>Declarative -> declarative -> action-oriented -> specific -> positive -> no-polite (HW)</i> 3. It will be safer if you use uncommon words (abstract) 4. It will be safer if you use both letters and numbers (concrete)
19 20 21	Statement 5&6&7	<i>Procedural -> imperative -> action-oriented -> specific -> positive -> no-polite (HW, H)</i> 5. Use uncommon words to make a good password (HW, abstract) 6. Use both letters and numbers to make a good password (HW, concrete) 7. Use uncommon words (H, abstract)
22	Statement 8	<i>Procedural -> imperative -> action-oriented -> specific -> positive -> yes-polite</i> 8. Please use uncommon words (H, abstract)
23	Statement 9	<i>Procedural -> imperative -> action-oriented -> specific -> negative -> no-polite</i> 9. Do not use common words (H, abstract)
<i>The suggestion in this case (suggestion in step 2) will be (1) abstract: add jokes. (2) concrete: add symbols.</i>		
24	Statement 1	<i>Declarative -> declarative -> password-oriented -> general -> positive -> no-polite</i> 1. Password is okay.
25 26	Statement 2&3	<i>Declarative -> declarative -> password-oriented -> specific -> positive -> no-polite (HW)</i> 2. You can improve your password by adding jokes (abstract) 3. You can improve your password by adding symbols (concrete)
27	Statement 4	<i>Procedural -> imperative -> action-oriented -> general -> positive -> no-polite</i> 4. Create a stronger password.

28	Statement 5	<i>Procedural -> imperative -> action-oriented -> general -> positive -> yes-polite</i> 5. Please create a stronger password.
29 30	Statement 6&7	<i>Procedural -> imperative -> action-oriented -> specific -> positive -> no-polite (HW)</i> 6. Add jokes to make your password stronger. (abstract) 7. Add symbols to make your password stronger. (concrete)
31	Statement 8	<i>Procedural -> imperative -> action-oriented -> specific -> positive -> yes-polite (HW)</i> 8. Please add jokes to make your password stronger. (abstract)
32	Statement 9	<i>Procedural -> imperative -> action-oriented -> specific -> negative -> no-polite (HW)</i> 9. Avoid passwords that are easy to guess
Group 3		
<i>The error message in this case (error message in step 2) will be</i>		
33	Statement 1	<i>Declarative -> declarative -> password-oriented -> general -> positive -> no-polite</i> 1. This is a very common password
34	Statement 2	<i>Declarative -> declarative -> password-oriented -> general -> negative -> no-polite</i> 2. Your password is weak
35	Statement 3	<i>Declarative -> declarative -> password-oriented -> specific -> negative -> no-polite</i> 3. Your password is too weak
<i>The requirement of the system in this case (policy in step 3) will be [a combination of uppercase, lowercase, and symbols]</i>		
36	Statement 1	<i>Declarative -> declarative -> password-oriented -> specific -> positive -> no-polite</i> The password needs to have a combination of uppercase letters, lowercase letters, and symbols
37	Statement 2	<i>Declarative -> modal -> password-oriented -> specific -> positive -> no-polite</i> The password should be a combination of uppercase letters, lowercase letters, and symbols
38	Statement 3	<i>Declarative -> modal -> password-oriented -> specific -> negative -> no-polite</i> The password should not be only uppercase letters, lowercase letters, or symbols
39	Statement 4	<i>Procedural -> imperative -> action-oriented -> specific -> positive -> no-polite</i> Use a combination of uppercase letters, lowercase letters, and symbols
40	Statement 5	<i>Procedural -> imperative -> action-oriented -> specific -> positive -> yes-polite</i> Please use a combination of uppercase letters, lowercase letters, and symbols
41	Statement 6	<i>Procedural -> imperative -> action-oriented -> specific -> negative -> no-polite</i> Do not use only uppercase letters, lowercase letters, or symbols
<i>The suggestion in this case (suggestion in step 3) will be [at least eight characters]</i>		
42	Statement 1	<i>Declarative -> declarative -> password-oriented -> specific -> positive -> no-polite</i> 1. Good passwords have at least eight characters (concrete)
43	Statement 2	<i>Procedural -> imperative -> action-oriented -> general -> positive -> no-polite</i> 2. Choose a more secure password
44	Statement 3	<i>Procedural -> imperative -> action-oriented -> general -> positive -> yes-polite</i> 3. Please choose a more secure password
45	Statement 4	<i>Procedural -> imperative -> action-oriented -> specific -> positive -> no-polite</i> 4. Try one with at least eight characters
<i>The error message in this case (error message in step 3) will be</i>		
46	Statement 1	<i>Declarative -> declarative -> password-oriented -> general -> negative -> yes-polite</i> 1. Sorry, your password is invalid
47	Statement 2	<i>Declarative -> declarative -> password-oriented -> general -> negative -> no-polite</i> 2. Your password is too easy to guess
48	Statement 3	<i>Declarative -> declarative -> password-oriented -> general -> positive -> no-polite</i> 3. Short passwords are easy to guess
49	Statement 4	<i>Declarative -> declarative -> password-oriented -> specific -> negative -> no-polite</i> 4. Your password is too short
50	Statement 5	<i>Declarative -> phrase only -> password-oriented -> general -> negative -> no-polite</i> 5. Invalid password